APPLICATION OF DIGITAL COMPUTERS

TO ON-LINE OPTIMISATION

J. MONK B.Sc.(Eng)

Thesis presented for

the Degree of Ph.D.

School of Engineering Science

The University of Warwick

England

December 1970

Abstract

This thesis is concerned with hill climbing systems which may be used for parameter optimisation. A simple system is described which uses the slope of the cost function to approach the zero slope condition corresponding to a unique optimum of a system characteristic.

A theoretical analysis using estimation theory methods is applied to a range of waveforms to assess their value in slope estimation in the presence of disturbances and system irregularities. The principal signals discussed are two and three level maximal-length sequences and square waves. Sine waves are used for comparison purposes. More complex optimisers using higher order models of the cost function are also considered with particular reference to the use of the special properties of three level maximal-length sequences. The results of the theoretical studies are confirmed by a series of experiments using an internal combustion engine test rig and a description of the preparation of the test rig is given. Hill climbing was carried out on the ignition angle to obtain maximum power at full throttle and trajectories for the hill climbing signals are shown.

Contents

CHAPTER 1

Introduction

The advent of cheap and reliable digital computers suitable for on-line process control has led to a new versatility in the application of control theory. In addition to conventional control, these computers enable complicated process optimisation schemes to be implemented with relative ease. The high working speed characteristic of computers makes it possible to complete complex modifications to the process data as it is received and then apply statistical principles to increase the accuracy of its analysis. Optimisers may be designed to modify process parameters in response to changes in operating conditions. In this way, the system parameters will be adapted to compensate for lack of knowledge, changes in process behaviour and environmental conditions. Further information about the process and the environment, which may not have been available for optimiser design, can be developed on-line by observing the process operation and then used to improve the system performance.

1.1 The Principles of Optimising Controllers

The three major procedures in an optimising controller are identification, decision and modification. Firstly, a model of the performance curves is developed by observing the process under control. The model may be extremely simple and only valid over a small range of operating conditions. In some cases it is possible to gather sufficient information for the model under normal operating conditions, but more generally, known probing signals are introduced to perturb the process and the results used in the identification of the system. The amplitude of the perturbation must be small to minimise any disturbance to the process.

An evaluation of the performance is then made and the decision procedure uses an adaptive algorithm or policy to locate possible improvements in performance. Normally, the procedure uses the current and possibly previous system models to evaluate a particular function of the process variables which in some way represents efficiency or cost. The maximum or minimum of this function will then represent the optimum performance. Most of the optimisation methods used to locate this optimum, such as conjugate gradient[31], require the value of the first derivative but this is not always essential[32].

Finally the modification procedure makes the necessary changes in the process parameters, providing these lie within the operating constraints and stability limitations of the process.

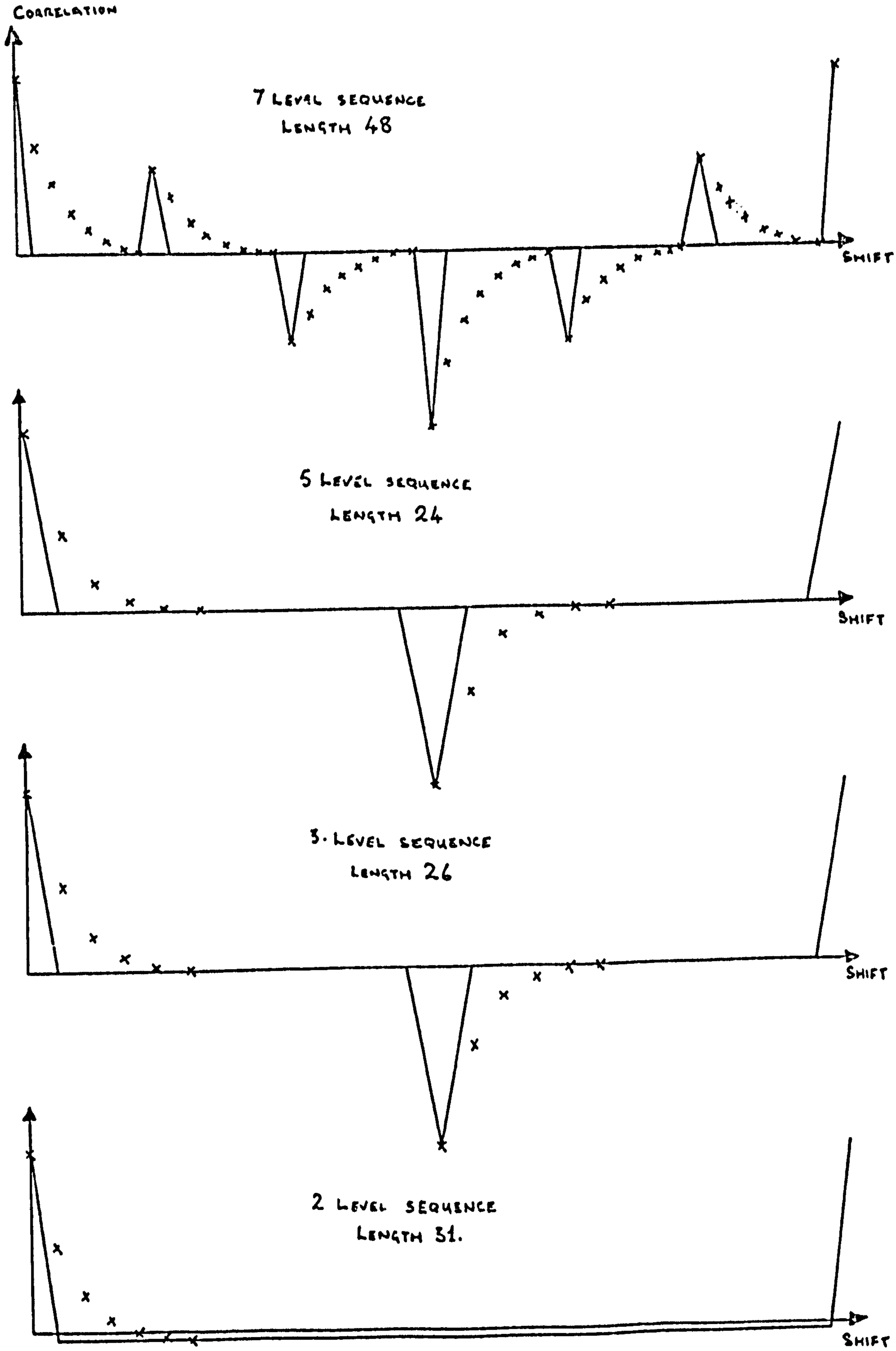
1.2 The Choice of Perturbations

Chapter 2 describes a simple optimiser which directly identifies the slope of the cost function at the operating point. This result is then proportional to the change in operating point taken to drive the system towards the maximum. Draper and Li[3] have obtained the estimate of the gain of the system being optimised, (which is the same as the first derivative of the cost function), by perturbing the input variable using a sinusoidal perturbation and phase detecting the response. Douce and King[29] introduced square waves as perturbations to simplify the detection problem. The necessary multiplication and integration is then reduced to a simple sign change and integration as the square wave has only two discrete levels. Roberts[28] showed that the sine wave is a preferable signal to square waves and random telegraph codes for a hill climber. In all these methods, the output of the phase detector is directly integrated and fed back to the input variable to achieve optimisation. Douce and Bond[30] have shown that the introduction of a sample and hold element suspending input variable adjustments eliminates parameter excursions and allows higher loop gains.

Van der Grinten[1] has proposed an alternative method using stochastic signals for the perturbation. Correlation techniques are used to give an estimate of the integral of the system impulse response which is proportional to the required change in parameter setting. He has also suggested using a binary equivalent of noise but major difficulties are encountered due to statistical fluctuations in the noise and finite experimental time errors. In addition, the delayed version of the noise used in the correlation may require

excessive storage. Douce and Ng[2] have proposed eliminating these errors and restrictions by using pseudo-random binary sequences in place of the truely random signals. These sequences are periodic but have an auto-correlation function similar to white noise. They are easily generated with shift registers[4] or the programmed equivalent so that delayed versions need not necessarily be stored.

Pseudo-random binary sequences are a sub-set of the class of maximal-length sequences. Sequences with a higher number of levels but similar auto-correlation functions may be generated using the same principles. The impulsive auto-correlation function implies that the perturbation of a linear system and correlation of its response gives the system impulse response directly[11] and that the system gain can be derived by integration for use in the hill climber. Fig. 1.2.1 shows the auto-correlation obtained with a first order system using two, three, five and seven-level sequences. For two and three level sequences, the correlation algorithm reduces to a simple sequence of additions and subtractions.



x SHOWS THE RESULT OF CROSS-CORRELATION

Fig. 1.2.1 Auto-correlation and results of cross-correlation experiments for 2, 3, 5 and 7 level maximal length sequences

1.3 Summary of Research

The research for this thesis was primarily concerned with the effect of the choice of perturbation signal on the performance of a single parameter optimisation scheme. Chapter 2 surveys and analyzes the use of a range of waveforms as perturbations for a simple optimiser with particular reference to implementation using a digital computer. The study is concerned with the effect on the system gain estimate of non-linearities, the selection of higher frequency perturbations and the presence of steady state levels, slow drifts and stochastic disturbances at the output of the system. Methods for the elimination of errors due to drifts and steady state levels are surveyed with particular reference to pseudo-random binary sequences. A matrix notation has been introduced into the analysis wherever possible as it provides results in a form directly applicable to digital algorithms.

Chapter 3 considers some of the alternative optimiser decision procedures which could be implemented. The effects of linearly filtering the estimates of system gain by introducing additional dynamic terms into the optimiser computing elements are investigated. Secondly, the two derivative hill climbers introduced by Godfrey and Clarke[27] are discussed. These utilise the special properties of pseudo-random ternary sequences which allow the simultaneous estimation of the first and second derivatives of a quadratic cost function[26] on a particular system configuration. A study is made of the alternative estimation schemes which could be used in this type of hill climber and some of the restrictions noted. The advantages and disadvantages of extending these principles to higher order cost function models are also discussed.

The theoretical results of these two chapters are then applied to a working process. Van der Grinten had carried out practical studies on a gas fired boiler[33] and Draper and Li[3] had obtained meaningful results using a special internal combustion rig. The most suitable process available proved to be an internal combustion engine test rig. Several new instruments were developed to achieve the full use of the potential speed of a digitally implemented optimiser, and a mathematical model of the test rig was built as an aid in designing the on-line controllers and the optimisers. Finally a complete suite of programmes was written to allow on-line running, control and optimisation of the engine test rig. This work is described in Chapter 4 and fig. 1.3.1 shows the fully instrumented engine test rig.

Enquiries within the motor industry showed that the adjustment of the ignition angle is currently based on a function of the speed and the inlet manifold pressure. The relationship between ignition advance and speed is found by running the engine at full throttle for a particular loading device setting and manually adjusting the ignition angle to give maximum power. This is then repeated for several values of load settings and a mechanical device designed to adjust the ignition setting according to the results.

An automatic experiment was implemented in which the engine could be held at constant speed while the throttle was opened and then the power maximised by adjusting the ignition setting. The relationship between the speed and the ignition angle to give maximum power was then established by repeating the experiment at a series of different speeds. The results of a series of these experiments using pseudo-random binary sequences, pseudo-random ternary sequences

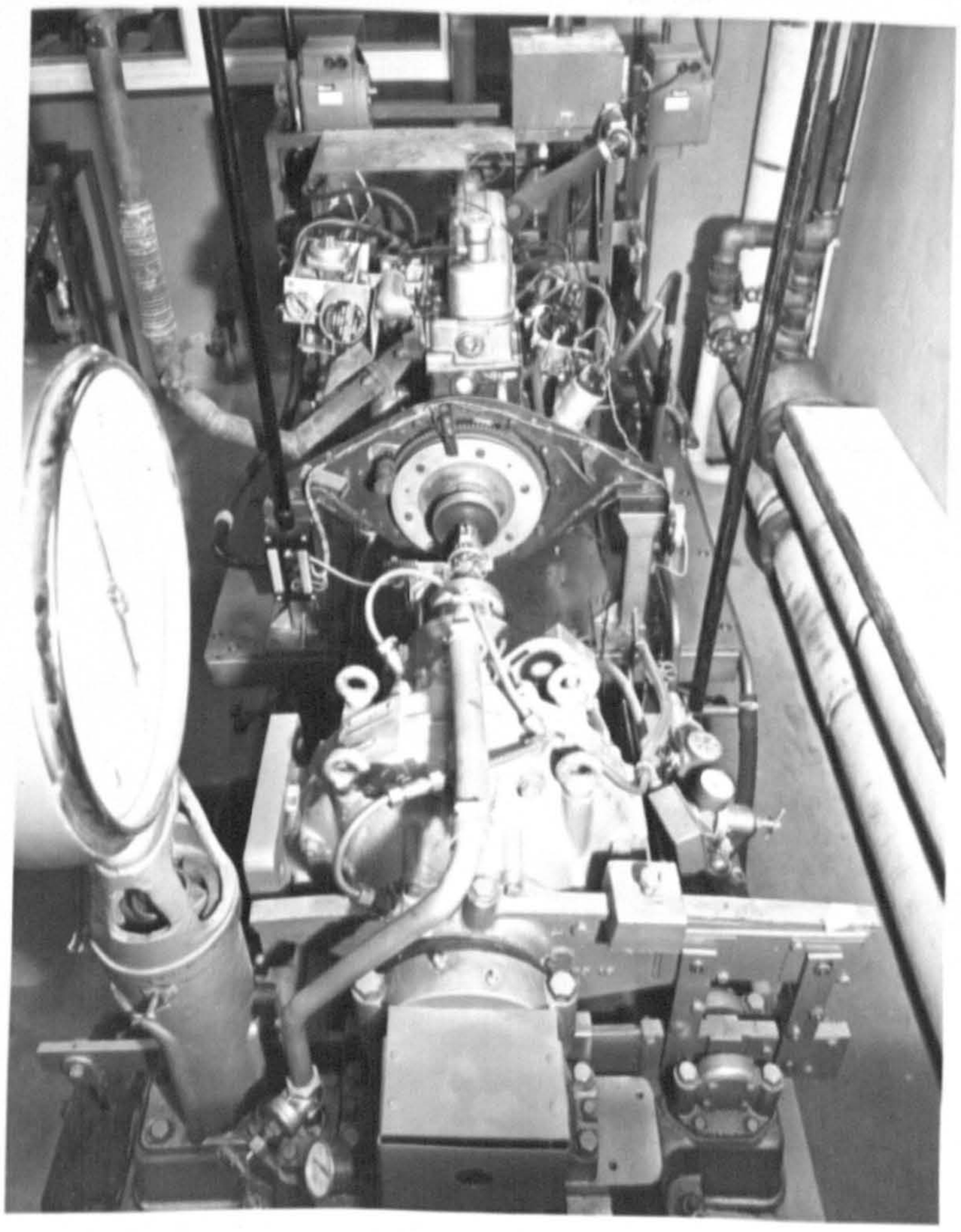


Fig. 1.3.1 The Fully Instrumented Engine Test Rig

and square wave perturbations with many different optimiser parameter settings are reported in Chapter 5.

The principal contributions of this thesis are the survey, unification and amplification of results concerning the application of specific waveforms - particularly pseudo-random sequences and square waves - in the identification of system parameters. This is achieved, where relevant by the application of matrix methods and results of estimation theory, the validity of these methods has been tested by applying the results to a working process.

CHAPTER 2

Basic Optimisers

Introduction

In this chapter, a dynamic model of the optimiser is developed in the form of a feedback control system and by making assumptions of linearity, conventional control theory is applied to derive stability criteria, transient performance and the effect of process disturbances on the system output. The model assumes that the slope of the cost function can be determined exactly in the absence of noise, but it is shown that the presence of low frequency drifts, non-linearities or the use of an incorrect perturbation frequency may introduce errors into this estimate. As these errors and those caused by process noise are all a function of the perturbation and the estimation method used, a range of specific perturbation signals are studied. The matrix methods previously applied by Clarke[9] to p.r.b.s. are extended to all the perturbations considered and provide results directly applicable to optimisers using digital computing elements.

2.1 Dynamic Performance of the Optimiser

2.1.1 Model of the Optimiser

A schematic diagram of a simple optimiser operating on a system is given in fig. 2.1.1. By introducing a small perturbation into the system and noting its effect at the output, an estimate of the slope of the cost function may be computed. All past values of the estimate of the gain of the system are summed to give the latest operating point and on receipt of a new estimate of gain, the operating point is incremented. If the estimator is used immediately after a change in operating point, the output will contain a transient component in addition to the response to the perturbation. A true value of the system gain will not be available until the transient has died away and the performance of the optimiser will be severely dependent on the detailed dynamic characteristics of the system. The optimisers discussed therefore commence estimation of the system gain when the system has settled after a change in operating point has occurred. The time interval between estimates is then not less than the sum of the minimum system settling time and the perturbation period. The model can therefore be treated as a sampled data system with a sampling time given by the period between successive estimates.

For analysis, the system dynamics, the cost function and the estimation procedure are replaced by a device which samples the operating point of the system and produces the gain estimate one sampling instant later. This device will be a time delay and a stationary, non-linear function given by the first derivative of the

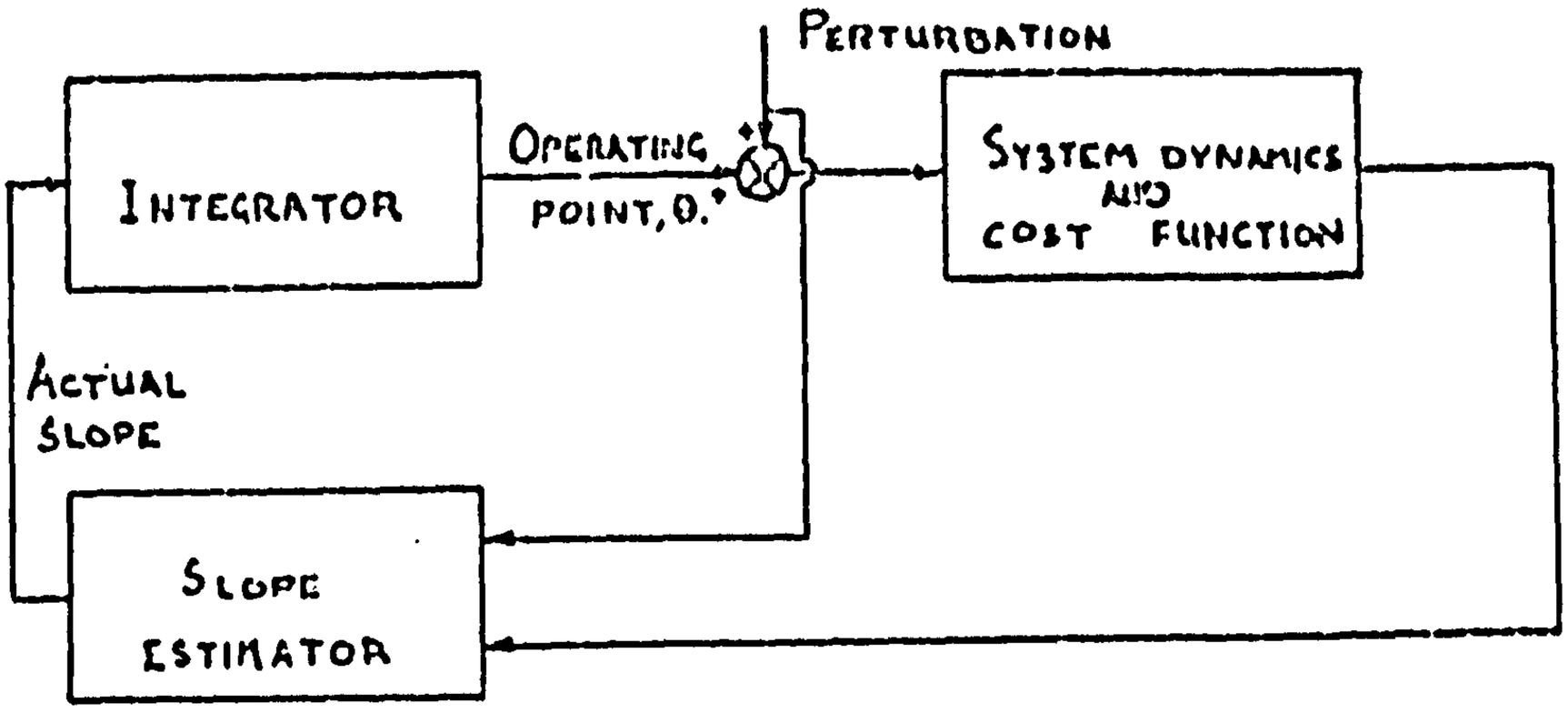


Fig. 2.1.1 - Schematic diagram of a simple optimiser

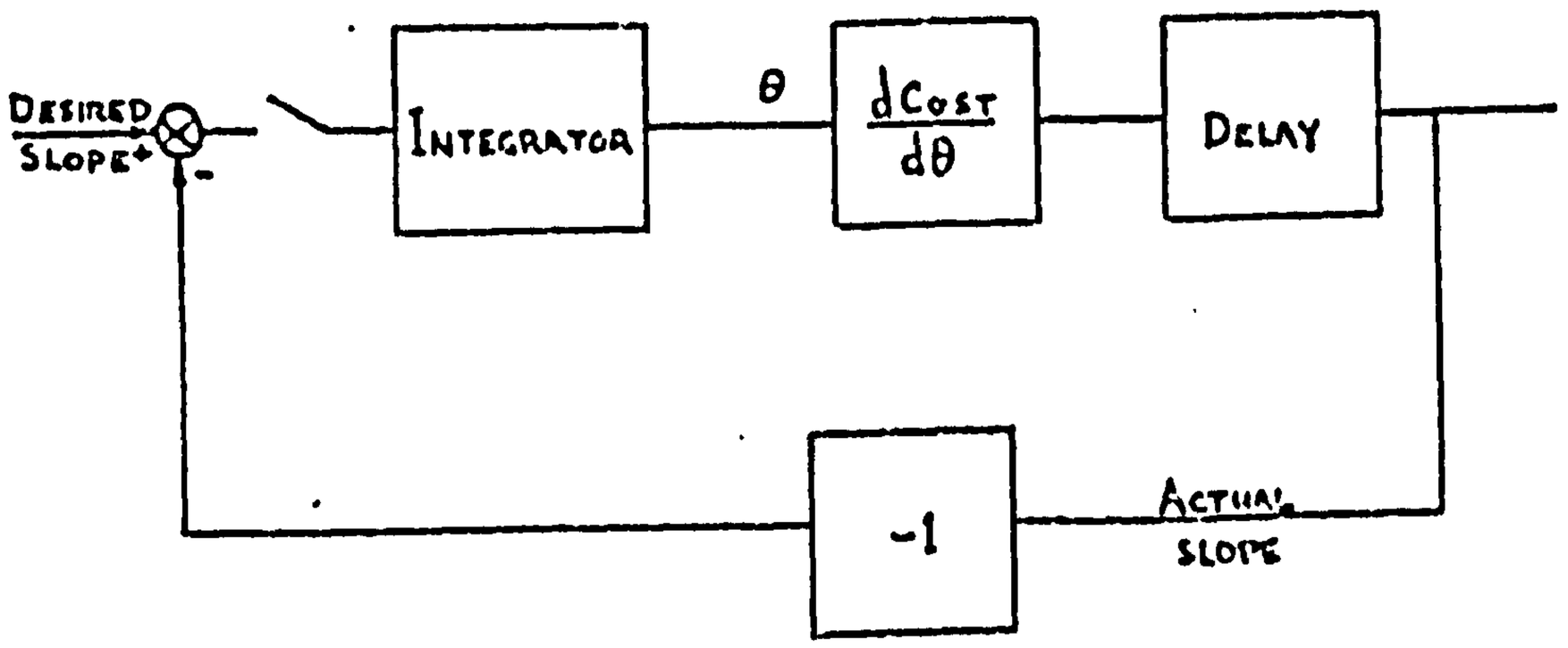


Fig. 2.1.2 - Optimiser as a control system

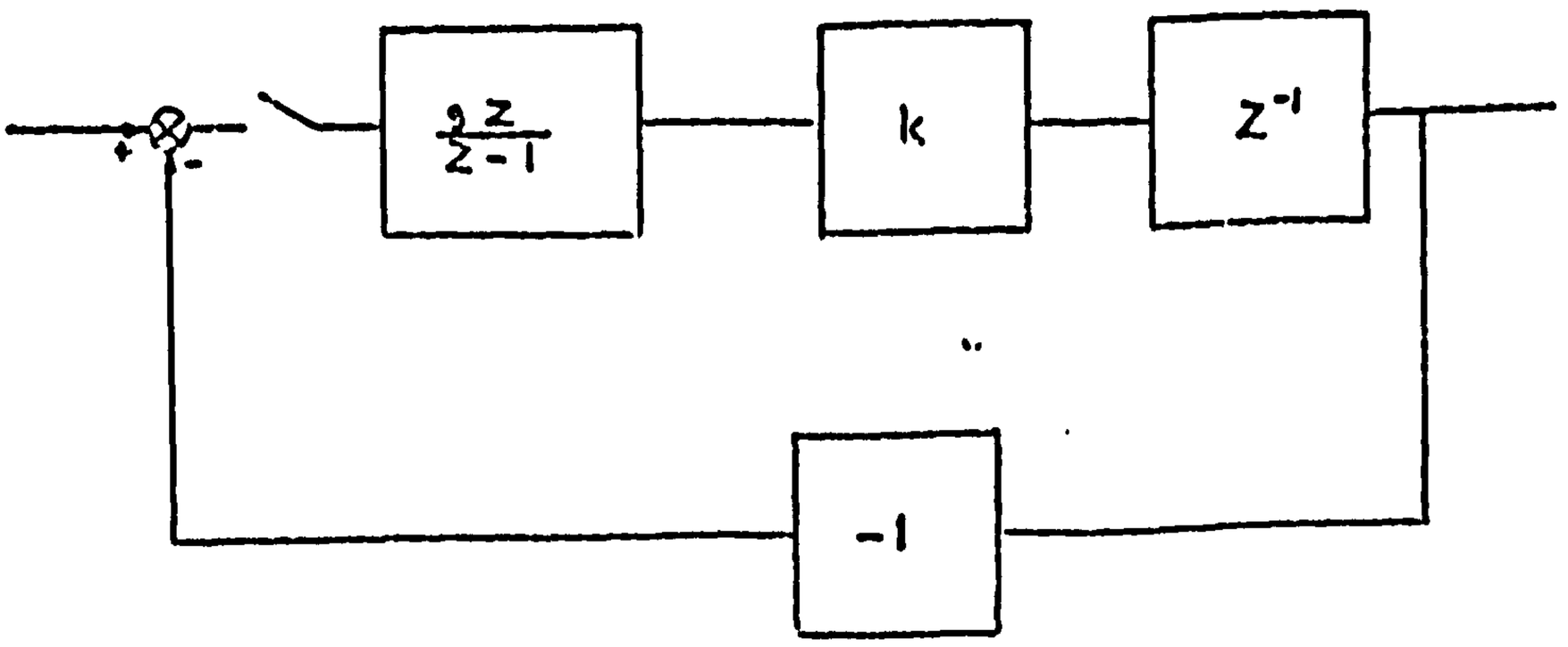


Fig. 2.1.3 - Z transform model of the optimiser

cost function. The model of the optimiser and the system is shown in fig. 2.1.2 and has been drawn as an ordinary closed loop control system to enable conventional control analysis to be applied. The input to the control system, (the desired slope), will normally be zero but may be used to temporarily disturb the system from the optimum to observe its transient response performance.

2.1.2 Dynamic Response

The analysis may be simplified by assuming a quadratic cost function and consequently a linear relationship between the estimated gain and the operating point. The slope of this transfer characteristic will be given by the second derivative of the cost function and will therefore be positive when searching for a minimum and negative when searching for a maximum. The linear z transform model of the optimiser is shown in fig. 2.1.3, where the integrator has gain g and the second derivative of the cost function is K, a constant. The forward path transfer function is given by

$$\frac{Kg}{(z - 1)}$$

and therefore the closed loop transfer function is

$$\frac{Kg}{z - (1 + Kg)}$$

which will be stable when the pole at $(1 + Kg)$ lies within the unit circle, that is

$$-2 < Kg < 0.$$

The dynamic performance for a range of operating points is shown in

fig. 2.1.4. When *Kg* equals *-1*, only one step is required to reach the optimum from any operating point.

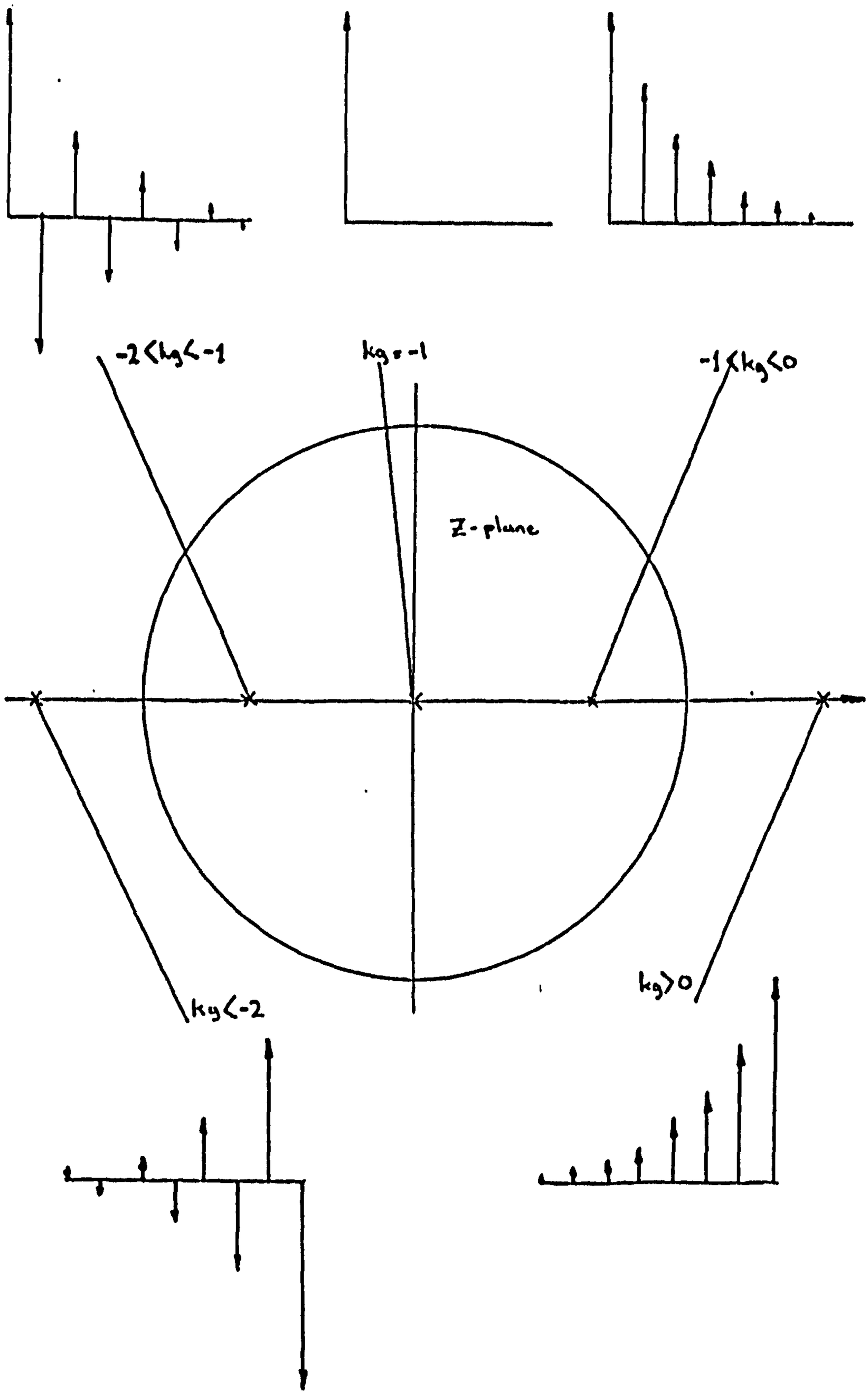


Fig 2.1.4 Dynamic performance of the optimiser.

2.2 Estimation for a Noise Free Systems

2.2.1 Basic Estimators

The sampled response $\langle y \rangle$ of a dynamic system perturbed by a sequence $\langle u \rangle$ is given by the convolution integral,

$$y(m\lambda) = \int_0^\infty h(t)u(m\lambda - t)dt,$$

where λ is the time between samples and $h(t)$ is the system impulse response.

If the generator of $\langle u \rangle$ holds its value between sampling intervals, then

$$y_m = \sum_{n=1}^{\infty} h_n u_{m-n},$$

where $h_n = \int_{(n-1)\lambda}^{n\lambda} h(t)dt$

and is termed the weighting sequence.

An optimiser requires the steady state gain of the system,which is given by the final value of the system output in response to a unit step. This is equivalent to

$$\int_0^\infty h(t)dt = \sum_{n=1}^{\infty} h_n = s, \text{ say.}$$

For most practical systems, the weighting sequence becomes negligible after a settling time. If $\langle u \rangle$ is periodic over the settling time

and this corresponds to N samples then

$$y_m = \sum_{n=1}^{N} h_n u_{m-n} \quad ,$$

which may be written in matrix notation as

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_N \end{bmatrix} = \begin{bmatrix} u_N & u_{N-1} & \cdot & \cdot & \cdot & u_1 \\ u_1 & u_N & \cdot & \cdot & \cdot & u_2 \\ \cdot & & & & & \cdot \\ \cdot & & & & & \cdot \\ \cdot & & & & & \cdot \\ u_{N-1} & & & & u_1 & u_N \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ \cdot \\ \cdot \\ \cdot \\ h_N \end{bmatrix}$$

or more compactly,

$$\underline{y} = \underline{U}\underline{h}. \qquad (2.2.1)$$

Solving for $\underline{h}$ assuming $\underline{U}$ to be of rank N,

$$\underline{h} = \underline{U}^{-1}\underline{y}$$

or alternatively premultiplying both sides by $\underline{U}'$,

$$\underline{U}'\underline{h} = \underline{U}'\underline{U}\underline{y}, \qquad (2.2.2)$$

and assuming $\underline{U}'\underline{U}$ is of rank N gives a matrix form of the Wiener Hopf equation

$$\underline{h} = [\underline{U}'\underline{U}]^{-1}\underline{U}'\underline{y}, \qquad (2.2.3)$$

where $\underline{U}'\underline{U}$ *is of the form*

$$N\begin{bmatrix} \phi_0 & \phi_1 & \phi_2 & \cdot & \cdot & \cdot & \phi_{N-1} \\ \phi_i & \phi_0 & & & & & \cdot \\ \phi_2 & & \cdot & & & & \cdot \\ \cdot & & & \cdot & & & \cdot \\ \cdot & & & & \cdot & & \\ \cdot & & & & & \phi_0 & \phi_1 \\ \phi_{N-1} & & & & & \phi_1 & \phi_0 \end{bmatrix},$$

and ϕ_m is the autocorrelation function of $\langle u \rangle$ for a shift of m and is given by,

$$\phi_m = \frac{1}{N} \sum_{n=1}^{N} u_n u_{m+n} \cdot \cdot$$

From the solution of these equations, the final value of the step response, s, is given by

$$s = \underline{1}'\underline{h},$$

where $\underline{1}$ represents a column vector with unit elements. The computation of s will be trivial when $\underline{U}'\underline{U}$ is a scalar matrix.

2.2.2 The Effect of Non-Zero Steady State System Output on Slope Estimation

Because of the non-linear cost function, variations in the parameter settings of the optimiser will alter the system gain and also the steady state output level. For an output level of a_0 over the period of perturbation,

$$\underline{y} = \underline{Uh} + a_0\underline{1}, \qquad (2.2.4)$$

and from this equation,

$$\underline{\hat{h}} = \underline{h} - \alpha_0 [\underline{U}'\underline{U}]^{-1}\underline{U}'\underline{1},$$

where the second term represents an error in the estimated dynamic characteristics.

It has been suggested[12] that for p.r.b.s., equation 2.2.4 may be made independent of α_0 by subtracting the mean of the sequence $\langle u \rangle$ before correlation. The matrix $\underline{U}'\underline{U}$ will then be of rank $N - 1$ however, and the solution must be written as

$$\underline{U}'\underline{U}\underline{h} = \underline{U}'\underline{y}.$$

This will be true for all waveforms with zero mean since $\underline{U}'\underline{1}$ is then zero.

As the original equation has only N degrees of freedom and there are now $N + 1$ parameters, additional information is necessary for a complete solution. One assumption which has been made[7] is that practical systems do not respond instantaneously, implying that h_1 is zero. The system equation may then be written as

$$\underline{y} = \underline{T}\underline{p} \qquad (2.2.5)$$

where

$$\underline{T} = \begin{bmatrix} 1 & u_{N-1} & u_{N-2} & \cdot & \cdot & \cdot & u_1 \\ 1 & u_N & u_{N-1} & \cdot & \cdot & \cdot & u_2 \\ 1 & u_1 & u_N & \cdot & \cdot & \cdot & u_3 \\ \cdot & \cdot & & & & & \cdot \\ \cdot & \cdot & & & & & \cdot \\ 1 & u_{N-2} & & & & u_1 & u_N \end{bmatrix} \quad \text{and} \quad \underline{p} = \begin{bmatrix} \alpha_0 \\ h_2 \\ h_3 \\ \cdot \\ \cdot \\ h_N \end{bmatrix}.$$

This may be solved to give $\underline{h}$ and then σ.

2.3 Effects of Noise

2.3.1 Effects of noise on optimiser performances

The effect of noise in the system being optimised is to introduce an uncertainty into the estimated value of the system gain. Consider equation 2.2.5, where the true value of $\underline{y}$ is now given by

$$\underline{y} = \underline{T}\underline{p} + \underline{e}, \qquad (2.3.1)$$

where $\underline{e}$ is a vector representing the noise at the output of the system, and the estimate of $\underline{p}$ is given by

$$\hat{\underline{p}} = [\underline{T}'\underline{T}]^{-1}\underline{T}'[\underline{T}\underline{p} + \underline{e}]$$

$$= \underline{p} + [\underline{T}'\underline{T}]^{-1}\underline{T}'\underline{e}$$

The covariance of the parameter $\hat{\underline{p}}$ will be given by

$$E([\hat{\underline{p}} - \underline{p}][\hat{\underline{p}} - \underline{p}]') = [\underline{T}'\underline{T}]^{-1}\underline{T}'E(\underline{e}\underline{e}')\underline{T}[\underline{T}'\underline{T}]^{-1} \quad .$$

If the elements of $\underline{e}$ are serially independent and of zero mean and variance σ^2,

$$E(\underline{e}\underline{e}') = \sigma^2\underline{I}_N$$

and $$covar(\hat{p}) = \sigma^2[T'T]^{-1} \qquad (2.3.2)$$

The uncertainty in $\hat{\underline{p}}$ may be incorporated into the optimiser model by adding noise to the estimated slope. Without loss of generality, this may be represented as a noise input to the system given in fig. 2.1.3 and its effect at the output may be determined from the relationship between the power spectral density of the noise, $\Phi_{ee}(\omega)$ and the power spectral density of the system output $\Phi_{00}(\omega)$, where,

$$\Phi_{00}(\omega) = \Phi_{ee}(\omega)\ |G(e^{j\omega T})|^2$$

where G is the z transfer function of the system and

τ is the time between slope estimates.

For the optimiser,

$$G(e^{j\omega\tau}) = \frac{kg}{e^{j\omega\tau} - (1 + kg)}$$

and therefore,

$$|G|^2 = \frac{k^2g^2}{(1 + kg)^2 + 1 - 2(1 + kg)\cos\omega\tau}$$

And the sampled error, Φ_{ee} will be assumed zero except for angular frequencies in the range $-\pi/\tau$ to π/τ. Provided the low frequency drifts in the system output are negligible or compensated for, the errors in successive inputs will be independent and Φ_{ee} will be constant at $\tau r^2/2\pi$, $-\pi/\tau < \omega < \pi/\tau$, where r^2 is the variance of the slope estimate.

Therefore,

$$\Phi_{oo}(\omega) = \begin{cases} \dfrac{\tau r^2k^2g^2/2\pi}{(1 + kg)^2 + 1 - 2(1 + kg)\cos\omega\tau}, & -\dfrac{\pi}{\tau} < \omega < \dfrac{\pi}{\tau} \\ 0, & \omega \geq \pi/\tau \\ & \omega \leq -\pi/\tau \end{cases}$$

and the total power at the output will be given by

$$\int_{-\pi/\tau}^{\pi/\tau} \Phi_{oo}(\omega)d\omega = 2\int_{0}^{\pi/\tau} \frac{(\tau r^2k^2g^2/2\pi)d\omega}{(1 + kg)^2 + 1 - 2(1 + kg)\cos\omega\tau}$$

Over the range of loop gains giving stability,

$$= \frac{r^2|kg|}{2 - |kg|} \qquad 0 < |kg| < 2$$

and the ratio of the variance of the operating point to the variance of the estimated slope for this range of loop gains is plotted in fig. 2.3.1.

2.3.2 Application of Least Squares to Estimation

In the previous section it was shown that the variance of the first derivative directly affects the *wander* of a hill climbing system and if the magnitude of this *wander* becomes intolerable, it will be necessary to reduce this variance. This may be done directly by taking several periods of the perturbation and averaging the outputs over these periods. The variance will then be reduced by $1/m$ for an experiment of m periods. Alternatively the period of the perturbation may be extended over a greater time, either by incrementing the sequence after several samples of output or by using a longer sequence of the same class. In equation *2.2.1* it was assumed that the settling time was equal to the perturbation period but if the perturbation is extended, the estimated weighting sequence becomes longer and it is possible to assume that the last few values of the weighting sequence are zero. The equation set will then be overdeterminate and, by using an alternative estimation technique, it is possible to use the spare equations to reduce the variance of the estimate.

If it is assumed that the system has settled after k ordinates of the weighting sequence, then

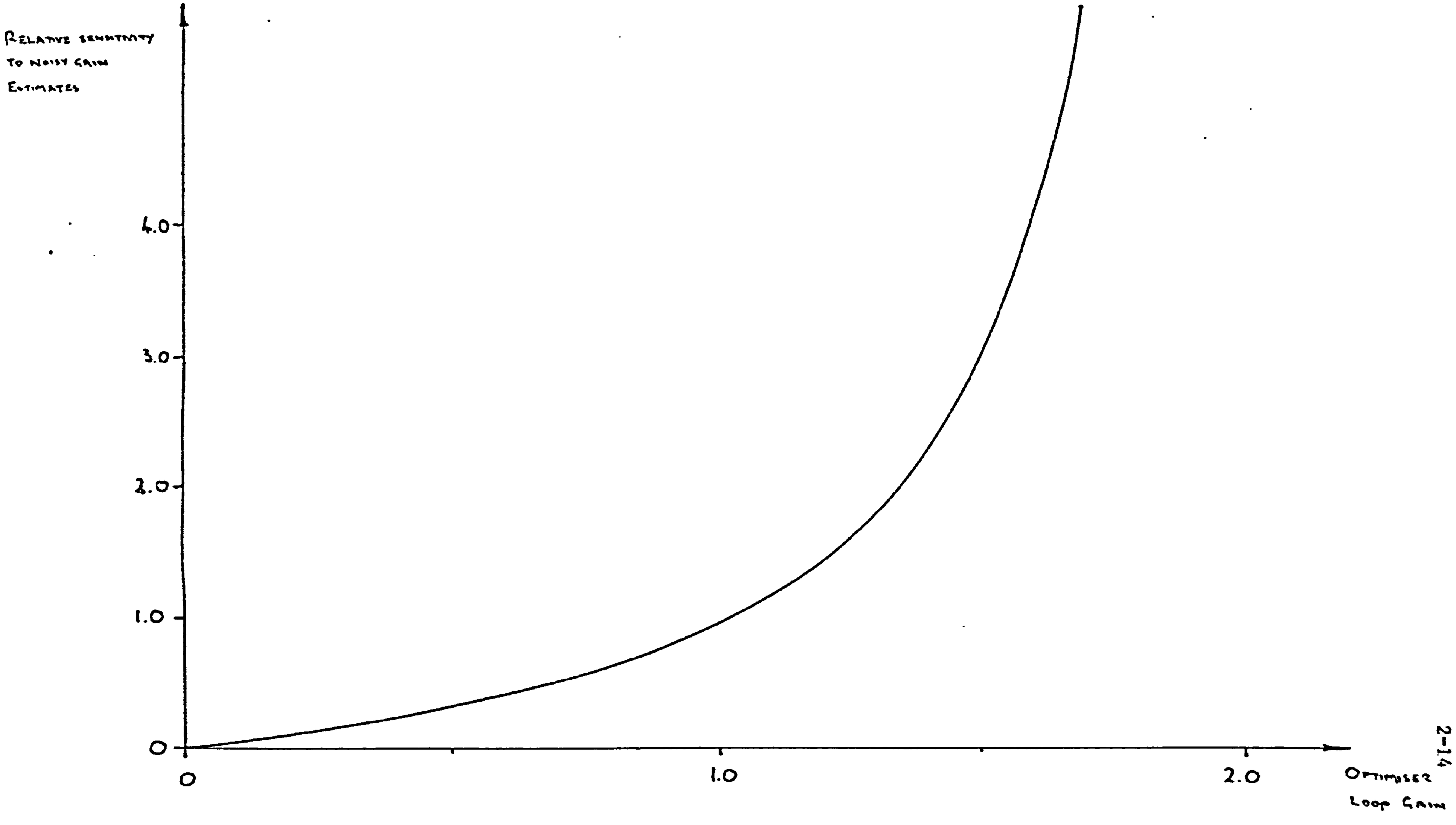


Fig 2.3.1. Effect of loop gain on optimiser noise sensitivity.

$$h_{k+1} = h_{k+2} = \ldots = h_n = 0,$$

and the system response may be written as

$$\underline{y} = \underline{X\theta} + \underline{e} \,, \qquad (2.3.3)$$

where

$$\underline{X} = \left[\begin{array}{cccccc|c} u_N & u_{N-1} & \cdot & \cdot & \cdot & u_{N-k+1} & 1 \\ u_1 & u_N & \cdot & \cdot & \cdot & u_{N-k+2} & 1 \\ \cdot & & & & & & \cdot \\ \cdot & & & & & & \cdot \\ \cdot & & & & & & \cdot \\ u_{N-1} & & & & & u_{N-k} & 1 \end{array}\right] \quad \text{and} \quad \underline{\theta} = \left[\begin{array}{c} h_1 \\ h_1 \\ \cdot \\ \cdot \\ \cdot \\ h_k \\ \hline a_0 \end{array}\right]$$

If the noise is serially independent, of zero mean and of variance σ^2, the principle of least squares[5] may be applied and

$$\underline{X'X\hat{\theta}} = \underline{X'y}$$

and providing $\underline{X'X}$ has rank k,

$$\underline{\hat{\theta}} = [\underline{X'X}]^{-1}\underline{X'y}$$

and $\quad covar(\underline{\hat{\theta}}) = \sigma^2[\underline{X'X}]^{-1}.$

The estimate of the step response s is given by

$$\hat{s} = [\underline{1}' \quad 0] \; \underline{\hat{\theta}} \,,$$

and its least squares estimate is therefore

$$\hat{s} = [\underline{1}' \quad 0][\underline{X'X}]^{-1}\underline{X'y} \qquad (2.3.4)$$

where $var(\hat{s}) = \sigma^2[\underline{1}' \quad 0][X'X]^{-1}\left[\begin{array}{c} \underline{1} \\ \hline 0 \end{array}\right]$ $\qquad (2.3.5)$

2.3.3 Structure of the $[\underline{X}'\underline{X}]$ matrix

The structure of the $[\underline{X}'\underline{X}]$ matrix is given by

$$\underline{X}'\underline{X} = N\left[\begin{array}{ccccccc|c} \phi_0 & \phi_1 & \phi_2 & \cdot & \cdot & \cdot & \phi_{k-1} & \\ \phi_1 & \phi_0 & & & & & \cdot & \\ \phi_2 & & \cdot & & & & \cdot & \\ \cdot & & & \cdot & & & \cdot & \bar{u}\,\underline{1} \\ \cdot & & & & \cdot & & \cdot & \\ \cdot & & & & & \phi_0 & \phi_1 & \\ \phi_{k-1} & \cdot & \cdot & \cdot & \cdot & \phi_1 & \phi_0 & \\ \hline \multicolumn{7}{c|}{\bar{u}\,\underline{1}'} & 1 \end{array}\right], \qquad (2.3.6)$$

where u is the mean value of sequence $\langle u \rangle$ given by

$$\bar{u} = \frac{1}{N} \sum_{n=1}^{N} u_n$$

and ϕ_m is the autocorrelation function for a shift of m and

$$\phi_m = \frac{1}{N} \sum_{n=1}^{N} u_n u_{m+n}.$$

2.3.4 Generalised Least Squares

When it is not possible to assume that the noise is serially independent, least squares becomes an inefficient estimator. If the auto-correlation function of the noise is known, generalised least squares may be applied, but as such results are specific to the noise present in the experiment, the application has not been developed further here.

2.3.5 Frequency Approach

If the effects of bias are excluded, the system response may be written as

$$\underline{y} = \underline{Uh} + \underline{e}.$$

Applying the discrete Fourier transform operator[10] $\underline{W}$,

$$\underline{Wy} = \underline{WUh} + \underline{We}, \tag{2.3.7}$$

where

$$\underline{W} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & w & w^2 & & w^{(N-1)} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 1 & w^{(N-1)} & w^{2(N-1)} & & w^{(N-1)(N-1)} \end{bmatrix},$$

and $w = e^{-2\pi j/N}$.

Now

$$\underline{WUh} = \underline{\Omega g},$$

where $\underline{\Omega}$ is the diagonal matrix whose elements are the frequency components of the perturbation and $\underline{g}$ is the frequency representation of the system. An estimate of the step response $\hat{s}$ will be given by

$$\hat{s} = \underline{1}'\underline{\hat{h}} = \underline{1}'\underline{W}^{-1}\underline{g}$$

$$= g_0,$$

the zero frequency response of the system.

When a frequency component does not appear in the perturbation

so that the correponding equation is absent in the system equations 2.3.7 or if any unknown bias term is present, further information about the system will be necessary for a complete solution. The conditions previously applied to supply further information were:

$$h_{k+1} = h_{k+2} = \ldots = h_{N} = 0$$

and in the frequency domain this becomes:

$$\begin{bmatrix} 1 & w^{k} & \cdot & \cdot & \cdot & w^{(N-1)k} \\ 1 & w^{k+1} & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ 1 & w^{N-1} & \cdot & \cdot & \cdot & w^{(N-1)(N-1)} \end{bmatrix} \underline{g} = 0 \qquad (2.3.8)$$

Any information still available after replacing the missing equations may be used to improve the variance of the estimates.

The equation set 2.3.8 may be cut to express a given set of k frequency components in terms of the remaining $N - k$ components. Let

$$\underline{g} = \left[\frac{\underline{g}^{*}}{\underline{g}^{**}}\right], \qquad \begin{array}{l} \text{where } g^{*} \text{ is a vector of length K} \\ \text{and } g^{**} \text{ is a vector of length N-k} \end{array}$$

where the set $\underline{g}^{*}$ is to be expressed in terms of the $\underline{g}^{**}$ components and let the corresponding re-arrangement of the truncated $\underline{W}$ matrix be $\underline{W_T}$ where

$$\underline{W_T} = [\underline{W_T}^{*} \quad \underline{W_T}^{**}]$$

and that of $\underline{\Omega}$ be

$$\underline{\Omega} = [\underline{\Omega}^{*} \quad \underline{\Omega}^{**}]$$

Then the re-arranged equation 2.3.8 will be

$$\underline{W_T}^{*}\underline{g}^{*} + \underline{W_T}^{**}\underline{g}^{**} = 0$$

and since $\underline{W_T}^{*}$ is square,

$$\underline{g}^{*} = - [\underline{W_T}^{*}]^{-1}\underline{W_T}^{**}\underline{g}^{**}$$

Equation 2.3.7 may then be written

$$\underline{Wy} = [\underline{\Omega}^{*} \quad \underline{\Omega}^{**}] \begin{bmatrix} -|\underline{W_T}^{*}|^{-1}\underline{W_T}^{**} \\ \hline \underline{I_k} \end{bmatrix} \underline{g}^{**} + \underline{We}$$

These equations may be solved using least squares when the error terms $\underline{e}$, and consequently the transformed error terms, are independent and of zero mean. As the quantities in these equations are complex, the application of least squares is not straightforward but if frequency components are known to be absent, the equations may be used to determine how much additional information will be required for a solution of the equation system in the time domain.

2.3.6 The Identification of Multivariable Systems

Consider a linear system with two inputs and one output where each input is stimulated by a different sequence, $\langle u \rangle$ and $\langle v \rangle$ respectively. The output $\langle y \rangle$ will be given by the sum of the separate outputs $\langle y_u + y_v \rangle$ and therefore,

$$\underline{y} = \underline{y_u} + \underline{y_v}$$

$$= \underline{Uh} + \underline{Vg} + a_0\underline{1} + \underline{e}$$

Assuming the settling times of the weighting sequences of the two paths $\underline{h}$ and $\underline{g}$ to be k_u and k_v respectively. The equation may be written,

$$\underline{y} = \underline{S\theta} + \underline{e},$$

where $\underline{\theta}' = [h_1 h_2 \ldots h_{ku} \quad g_1 g_2 \ldots g_{kv} \quad a_0]$

$$\underline{S} = [\underline{U_T} \quad \underline{V_T} \quad \underline{1}]$$

$$= \left[\begin{array}{cccc|cccc|c} u_N & u_{N-1} & \cdots & u_{N-ku+1} & v_N & v_{N-1} & \cdots & v_{N-kv+1} & 1 \\ \cdot & & & & \cdot & & & & 1 \\ \cdot & & & & \cdot & & & & \cdot \\ \cdot & & & & \cdot & & & & \cdot \\ u_1 & \cdot & \cdots & u_{N-ku+2} & v & \cdot & \cdots & v_{N-kv+2} & 1 \end{array}\right]$$

Applying least squares,

$$\underline{\hat{\theta}} = [\underline{S}'\underline{S}]^{-1}\underline{S}'\underline{y},$$

where

$$\underline{S}'\underline{S} = \left[\begin{array}{cc|c} \underline{\phi_{uu}} & \underline{\phi_{uv}} & \underline{1} \\ \underline{\phi_{vu}} & \underline{\phi_{vv}} & \\ \hline \underline{1} & & N \end{array}\right],$$

and $\underline{\phi_{uu}}$ and $\underline{\phi_{vv}}$ are the autocorrelation matrices for $\langle u \rangle$ and $\langle v \rangle$, and

$$\underline{\phi_{uv}} = \underline{\phi_{vu}}$$

is the cross-correlation matrix between $\langle u \rangle$ and $\langle v \rangle$.

The problem of inverting the $\underline{S}'\underline{S}$ matrix may be simplified by using uncorrelated sequences or using the same perturbation for both inputs but phase shifting the second $k_u + 1$ times with respect to the first. The cross-correlation matrix will be zero for the former and the latter, the four matrices in the top L.H. corner of $\underline{S}'\underline{S}$ will be replaced by the auto-correlation matrix of the sequence used for shifts up to $k_u + k_v - 1$.

2.4 Compensation for low frequency drift

The presence of low frequency drift of system output can cause errors in estimation and the methods applied for its removal in p.r.b.s. cross-correlation experiments are reviewed in this section. The methods are presented and extended using a matrix notation to generalise their solutions.

2.4.1 Use of Reference Phase

One of the earliest reports[11] of errors introduced into p.r.b.s. correlation experiments by ramp disturbances at the output of the system showed that a peak to peak error of the order of 15% occurred when the mean square value of the ramp was equal to the mean square value of the system impulse response. It was subsequently found[12] that when the bias was removed by using a sequence of zero mean, the cross-correlation between the ramp and the sequence became zero for one particular shift, termed the reference phase. This result was formalised[13] for the discrete case and later proved for all p.r.b.s.[14]

In matrix form,

$$\underline{y} = \underline{Uh} + a_0\underline{p}_0 + a_1\underline{p}_1,$$

where $\underline{p}_m' = [1^m 2^m \ldots N^m]$

and a_1 is the amplitude of the linear drift term.

If $\langle u \rangle$ is of zero mean, correlating gives:

$$\underline{U}'\underline{y} = \underline{U}'\underline{Uh} + a_1\underline{U}'\underline{p}_1$$

and $\underline{U}'\underline{Uh} = \underline{U}'\underline{y} - a_1\underline{U}'\underline{p}_1$,

where one element of $\underline{U}'\underline{p}_1$ will be zero when $\langle u \rangle$ is a p.r.b.s.

It was then shown[6] that the linear drift could be eliminated completely by using two periods of the p.r.b.s., starting at the reference phase, as an input and correlating one period of the sequence starting at the reference phase with successive blocks of output. The mth set of equations for blocks of N output values will be given by

$$\underline{y}_m = \underline{U}_m\underline{h} + (a_1\underline{p}_1 + m\underline{p}_0) + a_0\underline{p}_0,$$

where $\underline{U}_m$ is the mth to the $(m + N)$th rows of

$$\left[\begin{array}{c} \underline{U} \\ \hline \underline{U} \end{array}\right]$$

and the ramp disturbance is represented by $(a_1\underline{p}_1 + m\underline{p}_0)$. Correlating this equation with $\langle u_r \rangle$, the sequence commencing at the reference phase, the bias and drift terms will be zero and therefore,

$$\underline{U}_r'\underline{y}_m = \underline{U}_r'\underline{U}_m\underline{h},$$

where $\underline{U}_r$ is given by

$$\underline{U}_r = \begin{bmatrix} u_{r1} & u_{r2} & \cdots & u_{rN} & 0 & \cdots & 0 \\ 0 & u_{r_1} & & u_{r(N-1)} & u_{rN} & 0 & 0 \\ \vdots & & & & & & \\ 0 & \cdots & 0 & u_{r_1} & \cdots & & u_{rN} \end{bmatrix},$$

and the equation may be written:

$$\underline{U}_r'\underline{y} = \underline{U}_r'\left|\begin{array}{c} U \\ \hline U \end{array}\right|\underline{h} \equiv \underline{U}'\underline{Uh}.$$

2.4.2 Weighting Methods

Several authors[15, 16] have suggested weighting the correlation function and adding the results over several periods and their results have been generalised[17].

Consider the system equation for several periods of input,

$$\underline{y} = \begin{bmatrix} \underline{U} \\ \underline{U} \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} \underline{h} + \begin{bmatrix} \underline{Q}_1 \\ \underline{Q}_2 \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} \underline{a} = \begin{bmatrix} \underline{U} & \underline{Q}_1 \\ \underline{U} & \underline{Q}_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \end{bmatrix} \begin{bmatrix} \underline{h} \\ \underline{a} \end{bmatrix}$$

where $\underline{a}$ is the vector of polynomial coefficients and

$$\underline{Q}_m = \begin{bmatrix} 1 & nN - N + 1 & (nN - N + 1)^2 & \cdot & \cdot & \cdot & (nN - N + 1)^q \\ 1 & nN - N + 2 & \cdot & & & & \cdot \\ 1 & \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & \cdot & & & & \cdot \\ 1 & nN & nN^2 & & & & n^q N^q \end{bmatrix},$$

q being the highest order polynomial drift of interest. Weighting these equations,

$$\underline{Vy} = \underline{V} \begin{bmatrix} \underline{U} & \underline{Q}_1 \\ \underline{U} & \underline{Q}_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \end{bmatrix} \begin{bmatrix} \underline{h} \\ \underline{a} \end{bmatrix}. \qquad (2.4.1)$$

where $\underline{V}$ is the diagonal matrix whose elements form the weighting

coefficients and

$$\underline{V}'\underline{V} = \begin{bmatrix} \underline{V}_1 & \underline{0} & . & . & . & . & \underline{0} \\ \underline{0} & \underline{V}_2 & & & & & . \\ \underline{0} & \underline{0} & . & & & & . \\ . & . & & . & & & . \\ . & . & & & . & & . \\ . & . & & & & . & . \\ . & . & . & . & . & . & . \end{bmatrix} .$$

where $\underline{V}_m$ is an N square diagonal matrix.

Applying least squares to equation *2.4.1*,

$$\begin{bmatrix} \underline{U}' & \underline{U}' & . & . & . & . \\ \underline{Q}_1' & \underline{Q}_2' & . & . & . & . \end{bmatrix} \underline{V}'\underline{V}\underline{y} = \begin{bmatrix} \underline{U}' & \underline{U}' & . & . & . & . \\ \underline{Q}_1' & \underline{Q}_2' & & & & \end{bmatrix} \underline{V}'\underline{V} \begin{bmatrix} \underline{U} & \underline{Q}_1 \\ \underline{U} & \underline{Q}_2 \\ . & . \\ . & . \\ . & . \end{bmatrix} \begin{bmatrix} \underline{h} \\ \underline{a} \end{bmatrix}$$

(2.4.2)

and the R.H.S. may be written

$$\begin{bmatrix} \underline{U}'[\underline{V}_1 + \underline{V}_2 + \ldots]\underline{U} & [\underline{Q}_1'\underline{V}_1 + \underline{Q}_2'\underline{V}_2 + \ldots]\underline{U} \\ \underline{U}'[\underline{V}_1\underline{Q}_1 + \underline{V}_2\underline{Q}_2 + \ldots] & [\underline{Q}_1'\underline{V}_1\underline{Q}_1 + \underline{Q}_2'\underline{V}_2\underline{Q}_2 + \ldots] \end{bmatrix} \begin{bmatrix} \underline{h} \\ \underline{a} \end{bmatrix},$$

If $\underline{V}$ is chosen such that

$$\underline{U}'[\underline{V}_1\underline{Q}_1 + \underline{V}_2\underline{Q}_2 + \ldots] = 0,$$

then equation *2.4.2* may be cut into two independent sets of equations for $\underline{h}$ and $\underline{a}$ respectively. The former may be written,

$$[\underline{U}'\underline{U}' \ldots]\underline{V}'\underline{V}\underline{y} = [\underline{U}'[\underline{V}_1 + \underline{V}_2 + \ldots]\underline{U}]\ \underline{h},$$

and if $\underline{V}$ is further constrained to give

$$[\underline{V}_1 + \underline{V}_2 + \ldots] = \underline{I}_N,$$

then the estimate of $\underline{h}$ will be given by the usual cross-correlation but after the system response has been weighted by $\underline{V}'\underline{V}$.

Any noise on the system output will be transformed by the weighting coefficients and if the noise has been assumed to be homoscedastic, it will lose this property on transformation and the application of least squares may lead to an inefficient estimate of the system and drift parameters.

2.4.3 Ordinary Least Squares

For general polynomial drift, the system equation 2.3.1. will have the additional term $\underline{Q}_1\underline{a}$ and therefore over one period of the perturbation,

$$\underline{y} = [\underline{U}\underline{Q}_1] \begin{bmatrix} \underline{h} \\ \underline{a} \end{bmatrix} + \underline{e},$$

where a polynomial drift up to the qth order, $q \leq N - k$, may be eliminated. Applying least squares,

$$\begin{bmatrix} \hat{\underline{h}} \\ \hat{\underline{q}} \end{bmatrix} = \begin{bmatrix} \underline{U}'\underline{U} & \underline{U}'\underline{Q}_1 \\ \underline{Q}_1'\underline{U} & \underline{Q}_1'\underline{Q}_1 \end{bmatrix}^{-1} \begin{bmatrix} \underline{U}' \\ \underline{Q}_1' \end{bmatrix} \underline{y}.$$

Although this method will require considerable off-line computation, only one period of the perturbation is needed and the method applies to all waveforms.

2.5 Effect of non-linearities on identification

2.5.1 Analysis for a class of non-linear systems

The presence of a cost function within a system implies a non-linearity and this may introduce errors into the estimates of system dynamics and system gain. Consider the simplified system model shown in fig. 2.5.1 which assumes that the cost function and the dynamics are separable.

For the simplest system, the cost function will be a quadratic

$$a_0 + a_1 x + a_2 x^2,$$

where x is the instantaneous input to the cost function generator. The response of the system $y(t)$ to a sequence $\langle u \rangle$ with amplitude p will be given by:

$$y(t) = \int_0^\infty \{a_2 Z_1{}^2(t - \tau) + a_1 Z_1(t - \tau) + a_0\} h_2(\tau) d\tau,$$

where

$$Z_1(t - \tau) = \int_0^\infty \{\alpha + pu(t - \tau - z)\} h_1(z) dz$$

and α is the operating point. This may be written,

$$y(t) = \int_0^\infty \Big\{a_2\alpha^2 + a_1\alpha + a_0 + (a_1 + 2a_2\alpha) Z_2(t - \tau) + a_2 Z_2{}^2(t - \tau)\Big\} h_2(\tau) d\tau,$$

where

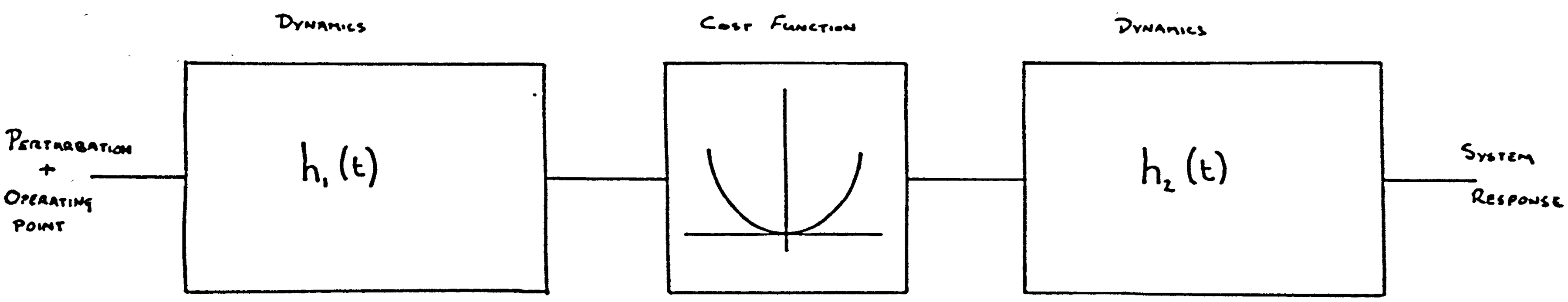


FIG 2.5.1. SIMPLIFIED SYSTEM STRUCTURE

$$Z_2(t - \tau) = \int_0^\infty pu(t - \tau - z)h_1(z)dz.$$

Correlating with the sequence, which has a clock interval of λ, and normalising with respect to the sequence amplitude p,

$$\phi_{uy}(m) = \left\{a_2\alpha^2 + a_1\alpha + a_0\right\}\frac{\bar{u}}{p}$$

$$+ \{a_1 + 2a_2\alpha\}\int_0^\infty h_2(\tau)\int_0^\infty h_1(z)\phi_{uu}(m\lambda - \tau - z)dzd\tau$$

$$+ pa_2\int_0^\infty h_2(\tau)\int_0^\infty h_1(z_1)\int_0^\infty h_1(z_2)\ .$$

$$\frac{1}{N\lambda}\sum_{N=1}^{N} u(n\lambda - \tau - z_1)u(n\lambda - \tau - z_2)u(n\lambda - m\lambda)dz_2dz_1d\tau$$

(2.5.1)

The first term is a constant and the impulse response of the system may be determined from the second when ϕ_{uu} is known. The error due to the non-linearity, represented by the third term, is proportional to the amplitude of the sequence and independent of the operating point. As the optimum is approached, the second term tends to zero and the final term may become dominant.

2.5.2 Effects of non-linearities in the frequency domain

The perturbation may be considered as the sum of several sine waves where the effect of passing through linear dynamics will be to alter the phase and amplitude relationship between harmonics. If the perturbation is then passed through a non-linearily, these will be further attenuated and phase shifted in the later dynamics.

When the overall response is correlated with the original waveform, the linear path will be identified but additional terms will appear at particular frequencies due to the intermodulation introduced by the non-linearity. If the frequency components of the perturbation sequence can be chosen so that these intermodulation frequencies do not exist in the original sequence, there will be no correlation between these additional terms and the sequence, and the error will be eliminated.

2.6 Effect of high frequency perturbations

The estimates of system dynamic response and gain have been obtained in previous sections assuming the settling time was known and well-defined. However, with an optimising system which continuously changes its operating point, it is possible that the system dynamics will change to give a longer settling time. The assumption will not then be valid and errors may be introduced into the estimates.

Consider a sequence of N samples perturbing a system which settles in q intervals but is assumed to settle in k intervals, where k is less than q. The true system output vector, in the absence of noise, will be given by:

$$\underline{y} = \left[\begin{array}{cccc|ccc} u_N & u_{N-1} & \cdots & u_{N-k+1} & u_{N-k} & \cdots & u_{N-q+1} \\ u_1 & u_N & \cdots & u_{N-k+2} & u_{N-k+1} & & \\ & & & & \vdots & & \\ & & & & \vdots & & \\ & & & & \vdots & & \\ & & & & \vdots & & \\ u_{N-1} & & & u_{N-k} & u_{N-k-1} & & \end{array}\right] \left[\begin{array}{c} h_1 \\ h_2 \\ \vdots \\ h_k \\ \hline h_{k+1} \\ \vdots \\ h_q \end{array}\right] + a_0\underline{1}$$

Using the notation of section 2.3.2,

$$\underline{y} = \underline{X\theta} + \left[\begin{array}{ccc} u_{N-k} & \cdots & u_{N-q+1} \\ u_{N-k+1} & & \\ \vdots & & \\ \vdots & & \\ u_{N-k-1} & & \end{array}\right] \left[\begin{array}{c} h_{k+1} \\ h_{k+2} \\ \vdots \\ \\ h_q \end{array}\right]$$

and using the result 2.3.4,

$$\hat{s} = [\underline{1}' \quad 0] \quad [\underline{X}'\underline{X}]^{-1}\underline{X}'\underline{X}\underline{0}$$

$$+ [\underline{1}' \quad 0] \; [\underline{X}'\underline{X}]^{-1}\underline{X}' \begin{bmatrix} u_{N-k} & \cdots & u_{N-q} \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ u_{N-k-1} & & \end{bmatrix} \begin{bmatrix} h_{k+1} \\ \cdot \\ \cdot \\ \cdot \\ h_q \end{bmatrix}$$

where the first term represents the true system gain and the second term is the error.

If q is an integral number of perturbation periods r, the error term may be written as

$$\varepsilon = [\underline{1}' \quad \underline{0}] \; [\underline{X}'\underline{X}]^{-1}\underline{X}' [\underline{U}_p \quad \underline{X}] \begin{bmatrix} \underline{\Sigma}_1 \\ \hline \underline{\Sigma}_2 \\ \hline 0 \end{bmatrix},$$

where

$$\underline{U}_p = \begin{bmatrix} u_{N-k} & \cdot & \cdot & u_1 \\ u_{N-k} & & & u_2 \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ u_N & & & \cdot \\ u & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ u_{N-k-1} & & & u_N \end{bmatrix}$$

and the nth elements of $\underline{\Sigma}_1$ and $\underline{\Sigma}_2$ are given by

$$\sum_{m=0}^{r-1} h_{k+mN+n} \text{ and } \sum_{m=1}^{r-1} h_{mN+n},$$

respectively. Then,

$$\varepsilon = \underline{1}'\underline{\Sigma}_2 + [\underline{1}' \quad \underline{0}]\ [\underline{X}'\underline{X}]^{-1}\underline{X}'\underline{U}_p\underline{\Sigma}_1$$

where the first term is independent of the waveform and $\underline{X}'\underline{U}_p$ may be further simplified to

$$N\begin{bmatrix} \phi_k & \phi_{k+1} & \phi_{k+2} & & \phi_{N-1} \\ \phi_{k-1} & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ \phi_2 & & & & \cdot \\ \phi_1 & \phi_2 & & & \phi_{N-k} \\ \hline \multicolumn{5}{c}{\bar{u}\ \underline{1}'} \end{bmatrix}$$

2.7 Application of Particular Waveforms

Careful design of an optimiser should lead to a system where the perturbation is of a sufficiently low frequency to eliminate errors caused by too short a perturbation period. If drift or bias is known to be present at the system output, it may be compensated for, thus removing these sources of error. The effects of noise and non-linearities cannot generally be eliminated and each class of waveform must be examined to determine the magnitude of the error.

2.7.1 Application using pseudo-random binary sequences

For a p.r.b.s. with unit clock rate and unit height,

$$\bar{u} = \frac{1}{N} \text{ and } \phi_m = \begin{cases} 1, & m = 0 \\ -\frac{1}{N}, & m \neq 0. \end{cases}$$

Therefore, substituting in equation 2.3.6,

$$\underline{X}'\underline{X} = \left[\begin{array}{c|c} (N+1)\underline{I}_k - \underline{J}_k & \underline{1} \\ \hline \underline{1}' & N \end{array}\right]$$

and using the results of A1.3,

$$[\underline{X}'\underline{X}]^{-1} = \frac{1}{(N+1)(N-k)} \left[\begin{array}{c|c} (N-k)\underline{I}_k + \underline{J}_k & -\underline{1} \\ \hline -\underline{1}' & N+1-k \end{array}\right] \qquad (2.7.1)$$

The estimates of the step response from equation 2.3.4 and the variance from equation 2.3.5,

$$\hat{s} = \frac{N}{(N+1)(N-k)} \sum_{n=1}^{N} y_n \left[\sum_{p=1}^{k} u_{n-p} - \frac{k}{N} \right]$$

and $$\text{var}(\hat{s}) = \frac{N k_\sigma^2}{(N+1)(N-k)}$$

The effects of the ratio of settling time to perturbation period on the variance of the estimate for a p.r.b.s. as shown in figs. 2.7.1 and 2.7.2.

Clarke[9] has shown that the impulse response obtained by this least squares estimator is equivalent to correlating the response of the system with the perturbation, and estimating the bias on the impulse response by averaging over the last $N - k$ ordinates. The weighting sequence is assumed to be zero after the system settling time of k elements.

2.7.2 P.r.b.s. in non-linear systems

When the estimation technique developed in the last section is applied to the correlation obtained in equation *2.5.1*, the estimate of the system gain will also contain an error term due to non-linear effects which in general will be non-zero. An optimiser using this sequence will therefore settle off the optimum where the error term is equal but of opposite sign to the system gain. If h_1 is a pure time delay, the error term is a constant and may be removed by testing it like the bias term. However, the effects of the non-linearity may be removed completely by performing several different experiments[18]. The effect of a quadratic non-linearity of a system composed of two cascaded first order lags at different operating points is shown in fig. 2.7.3.

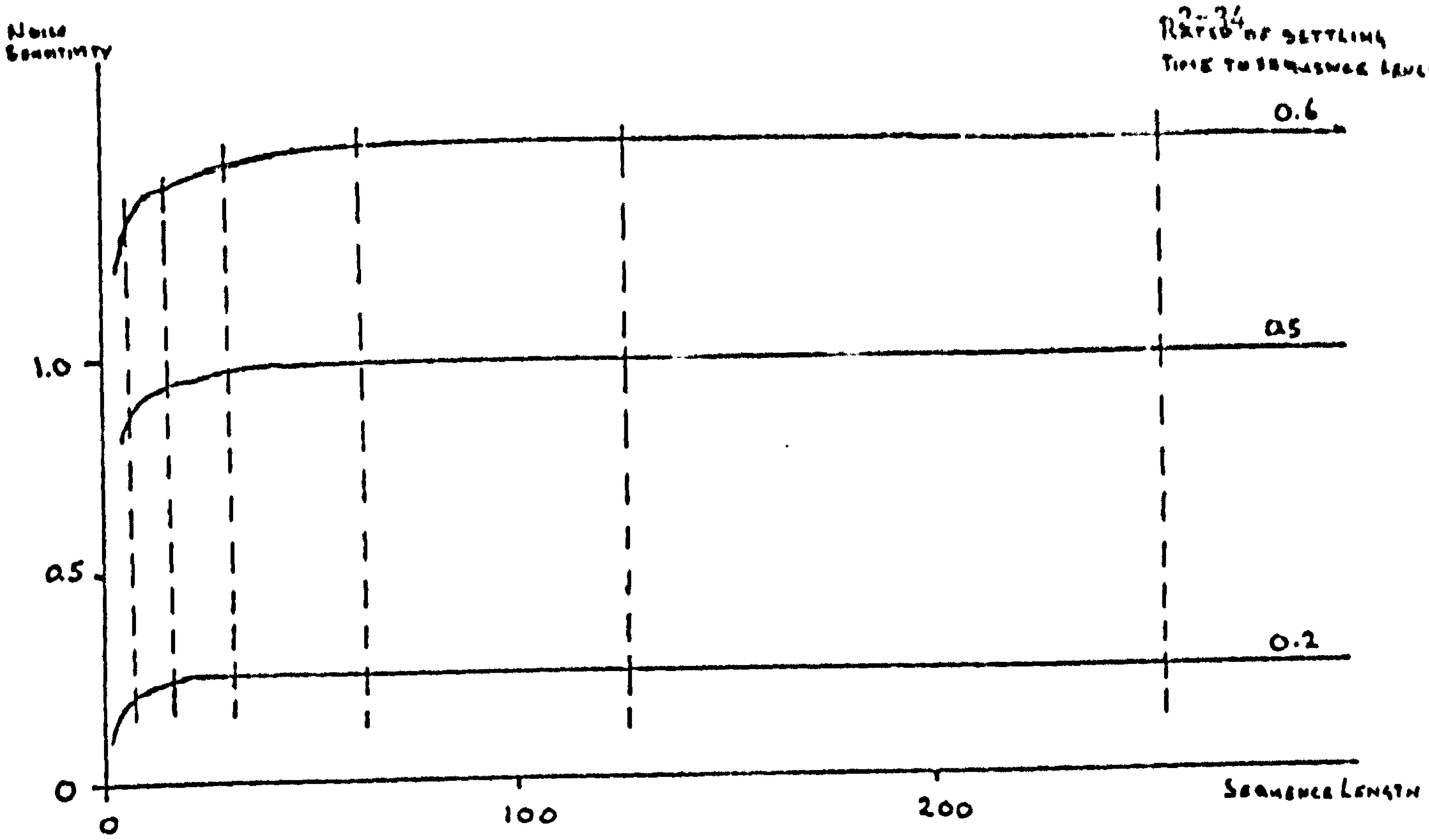


Fig. 2.7.1 - Effect of p.r.b.s. length on noise sensitivity of gain estimator

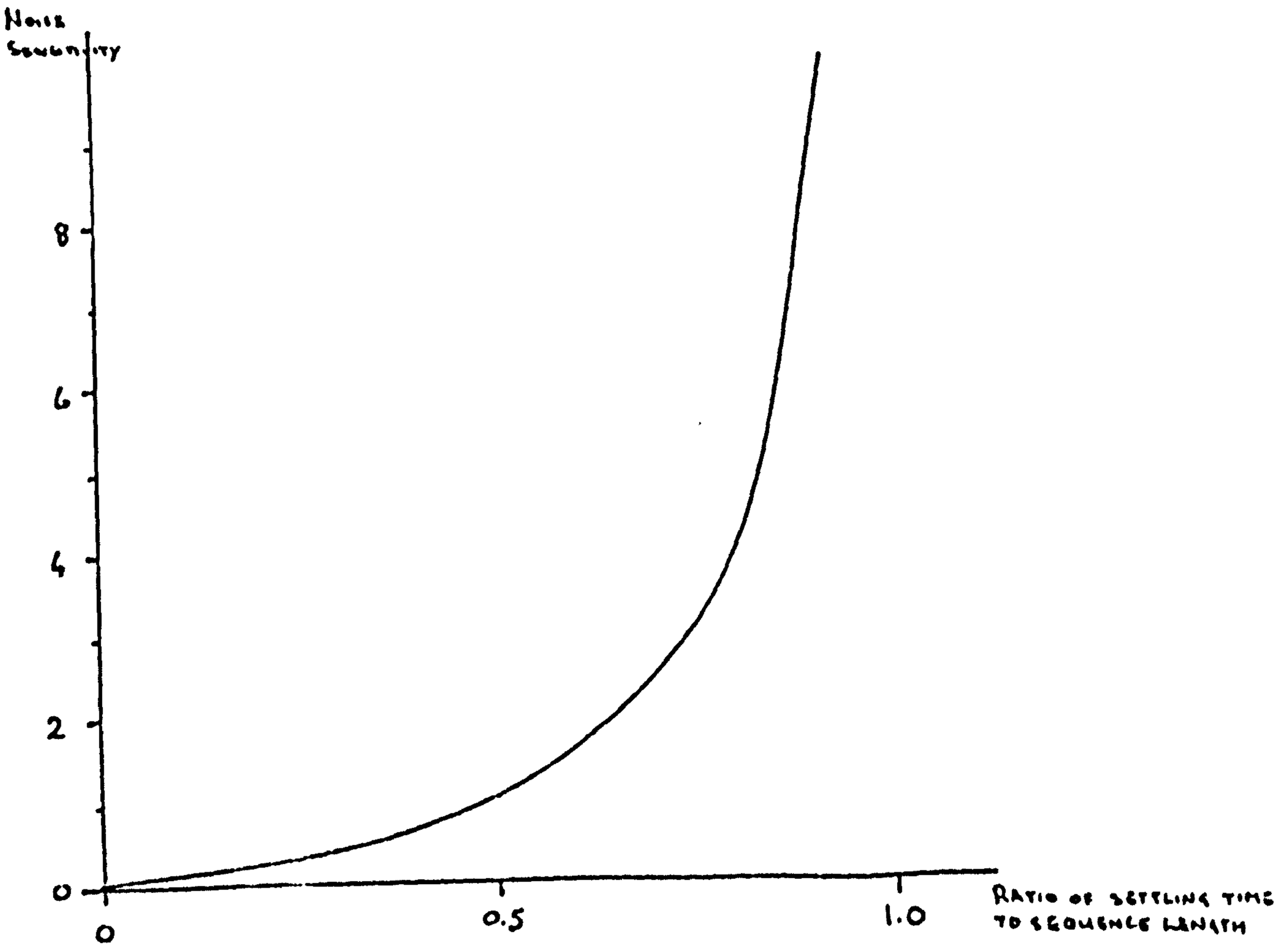


Fig. 2.7.2 - Effect of settling time on noise sensitivity of gain estimator using long p.r.b.s.

(a) (b) (c)

Fig. 2.7.3 - Effect of a quadratic nonlinearity on impulse response estimate

Note that the vertical scales are enlarged for (b) and further for (c)

(c) is closest to the optimum and (a) the furthest

2.7.3 Impulsive autocorrelation function with added bias

Consider a sequence $\langle c \rangle$ formed by adding a constant β to a sequence $\langle u \rangle$ with zero mean and impulsive autocorrelation function. For unit power, the autocorrelation matrix for $\langle u \rangle$ will be given by

$$\frac{N}{(N-1)}\,[N\underline{I}_N - \underline{J}_N].$$

Now $\phi_{cc} = \phi_{uu} + \beta^2$

and $\bar{c} = \beta$

therefore using the result of 2.3.6,

$$\underline{X}'\underline{X} = N \left[\begin{array}{c|c} \frac{1}{(N-1)}\left[N\underline{I}_k + \left((N-1)\beta^2 - 1\right)\underline{J}_k\right] & \beta\underline{1} \\ \hline \beta\underline{1}' & 1 \end{array}\right]$$

Inverting using A1.3,

$$[\underline{X}'\underline{X}]^{-1} = \frac{N-1}{N^2(N-k)} \left[\begin{array}{c|c} (N-k)\underline{I}_k + \underline{J}_k & -N\beta\underline{1} \\ \hline -N\beta\underline{1}' & \frac{N}{(N-1)}\left[N + ((N-1)\beta^2 - 1)k\right] \end{array}\right]$$

and applying equation 2.3.5,

$$\text{var}(\hat{s}) = \frac{(N-1)k\sigma^2}{N(N-k)}$$

If $\langle c \rangle$ is of unit power, this becomes,

$$\text{var}(\hat{s}) = \frac{(N-1)k\sigma^2}{N(N-k)(1-\beta^2)} \qquad (2.7.2)$$

Fig. 2.7.4 shows the effect on the variance of $\hat{s}$ for a range of settling times, and sequence lengths. Note that the result 2.7.2 is applicable to a p.r.b.s. with limits ± 1, for $\beta = 1/N$.

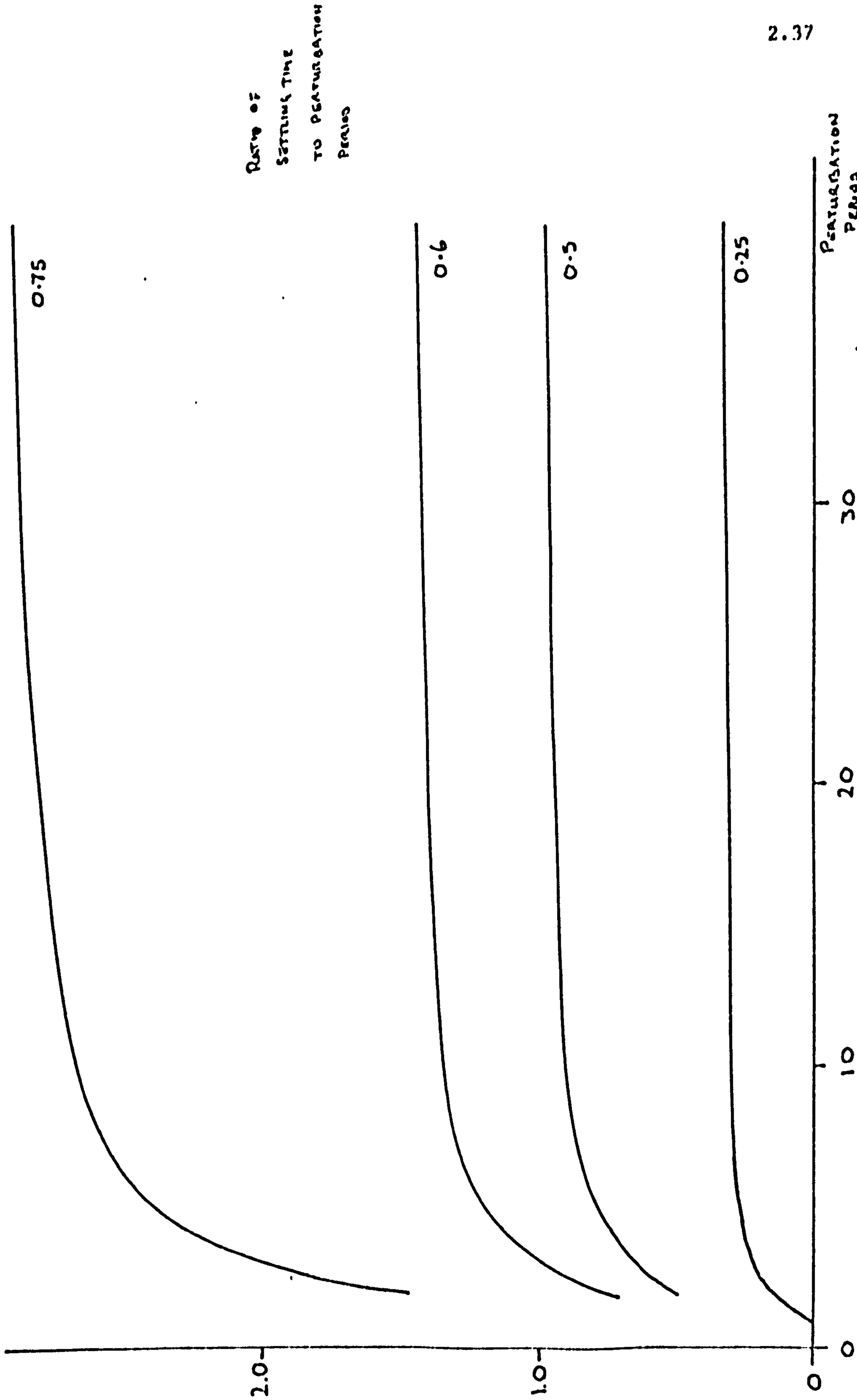

Ratio of Settling time to perturbation period
0.75
0.6
0.5
0.25
Perturbation Period
e estimator
30
20
10
0
period on the noise sensitivity of the
- Effect of perturbation p
2.0
1.0
0
Noise Sensitivity

Fig. 2.7.4

2.7.4 Note on Rotation Symmetry

Rotation symmetry of a waveform occurs when alternate half cycles are identical in shape but reversed in sign. Signals possessing this property have zero mean and contain only odd harmonics.

When the test signal has rotation symmetry, the matrix $\underline{X}'\underline{X}$ will have the slightly simpler form:

$$\underline{X}'\underline{X} = \left[\begin{array}{c|c} \underline{U}_T'\underline{U}_T & \underline{0} \\ \hline \underline{0}' & N \end{array}\right], \qquad (2.7.3)$$

where $\underline{U}_T$ is the $\underline{U}$ matrix truncated to the first k columns. The inverse of the matrix 2.7.3 may be found by inverting the upper and lower square matrices independently and therefore,

$$\hat{a}_0 = \frac{1}{N} \sum_{N=1}^{N} y_n,$$

and $$\mathrm{var}(\hat{a}_0) = \frac{\sigma^2}{N}$$

for all sequences possessing rotational symmetry. The matrix $\underline{U}_T'\underline{U}_T$ will be of rank $N/2$ and therefore the settling time of the system must be less than half the perturbation period to obtain a solution.

Two waveforms will be uncorrelated when they have no common frequency components, and therefore a sequence with rotation symmetry, and consequently no even harmonics, will be uncorrelated with any waveform whose fundamental is twice its own. If the second waveform also has rotation symmetry, the process may be extended to produce more uncorrelated waveforms. This property may be used in the iden-

tification of multivariable systems (2.3.6), but each time the fundamental frequency is doubled, the maximum settling time which may be identified is halved and experimental times may become excessive, if the time constants of the system are of the same order. Methods for generating sequences with rotation symmetry, impulsive autocorrelation functions and doubling of the fundamental frequency have been given by Briggs and Godfrey[19].

For the simple non-linear system described in section 2.5.1, the error term may be eliminated by using a sequence with rotation symmetry, as the odd harmonics will produce sums and differences of frequencies at the even harmonic frequencies when passed through an even order non-linearity, so that none of the intermodulation will correlate with the original sequence.

2.7.5 Three level m-sequences

A three level m-sequence possesses rotation symmetry and with levels ± 1 and 0, a power of $2(N + 1)/3$. Therefore

$$\underline{X}'\underline{X} = \begin{bmatrix} \frac{2(N+1)}{3}\underline{I}_k & \underline{0} \\ \underline{0}' & N \end{bmatrix}$$

where $k \leqslant N/2$ for a practical solution.

$$[\underline{X}'\underline{X}]^{-1} = \begin{bmatrix} \frac{3}{2(N+1)}\underline{I}k & \underline{0} \\ \underline{0}' & \frac{1}{N} \end{bmatrix}$$

Using equations 2.3.4 and 2.3.5,

$$\hat{s} = \frac{3}{2(N+1)} \sum_{m=1}^{N} y_m \sum_{n=1}^{k} u_{m-n}$$

and $\quad \mathrm{var}(\hat{s}) = \dfrac{3k\sigma^2}{2(N+1)}$

or for a sequence with unit power,

$$\mathrm{var}(\hat{s}) = \frac{k\sigma^2}{N}$$

Squaring a three level m-sequence gives a two level sequence of period $N/2$ and an impulsive autocorrelation function. The analysis for this waveform is covered in section 2.3.5 where $\beta = \dfrac{N+4}{3N}$ when the levels are made ± 1.

2.7.6 Square Wave Perturbations

The N-square circulant matrix $\underline{U}$ for a square wave perturbation with levels ± 1 is given by:

$$\underline{U} = \begin{bmatrix} -1 & 1 & 1 & \ldots & 1 & 1 & 1 & -1 & \ldots & -1 & -1 \\ -1 & -1 & 1 & \ldots & 1 & 1 & 1 & 1 & \ldots & -1 & -1 \\ \vdots & \vdots & & & & & & & & & \\ -1 & -1 & -1 & \ldots & -1 & -1 & 1 & 1 & \ldots & 1 & 1 \\ 1 & -1 & -1 & \ldots & -1 & -1 & -1 & 1 & \ldots & 1 & 1 \\ \vdots & & & & & & & & & & \\ 1 & 1 & 1 & \ldots & 1 & -1 & -1 & -1 & \ldots & -1 & 1 \\ 1 & 1 & 1 & \ldots & 1 & 1 & -1 & -1 & \ldots & -1 & -1 \end{bmatrix}$$

and since the square wave possesses rotation symmetry, it may be treated as in section 2.7.4. Now,

$$\underline{U_T}'\underline{U_T} = \begin{bmatrix} N & N-4 & N-8 & . & . & . & N-4(k-1) \\ N-4 & N & & & & & . \\ N-8 & & . & & & & . \\ & & & & . & & . \\ & & & & & N & N-4 \\ N-4(k-1) & . & . & . & .. & N-4 & N \end{bmatrix}$$

and using the result of A14 and equations *2.3.4* and *3.3.5*

$$\hat{s} = \frac{1}{(N - 2k + 2)} \left\{ \sum_{n=\frac{N}{2}+k}^{N} y_n - \sum_{n=k}^{\frac{N}{2}} \right\}$$

and $\quad \text{var}(\hat{s}) = \dfrac{\sigma^2}{(N - 2k + 2)}$

which is plotted in fig. 2.7.5 for a range of sequence lengths and settling times.

2.7.7 Sine Wave Perturbations

The N-square circulant matrix $\underline{U}$ for a sine wave with unit amplitude is given by:

$$\underline{U} = \frac{1}{2j} \begin{bmatrix} \omega - \omega^{-1} & \omega^2 - \omega^{-2} & . & . & . & . & \omega^N - \omega^{-N} \\ \omega^N - \omega^{-N} & \omega - \omega^{-1} & & & & & \omega^{(N-1)} - \omega^{-(N-1)} \\ . & & & & & & . \\ . & & & & & & . \\ . & & & & & & . \\ . & & & & & & . \\ \omega^2 - \omega^{-2} & & & & & \omega^N - \omega^{-N} & \omega - \omega^{-1} \end{bmatrix}$$

where $\omega = e^{\frac{2\pi j}{N}}$, and therefore the matrix $\underline{U_T}'\underline{U_T}$ is given by:

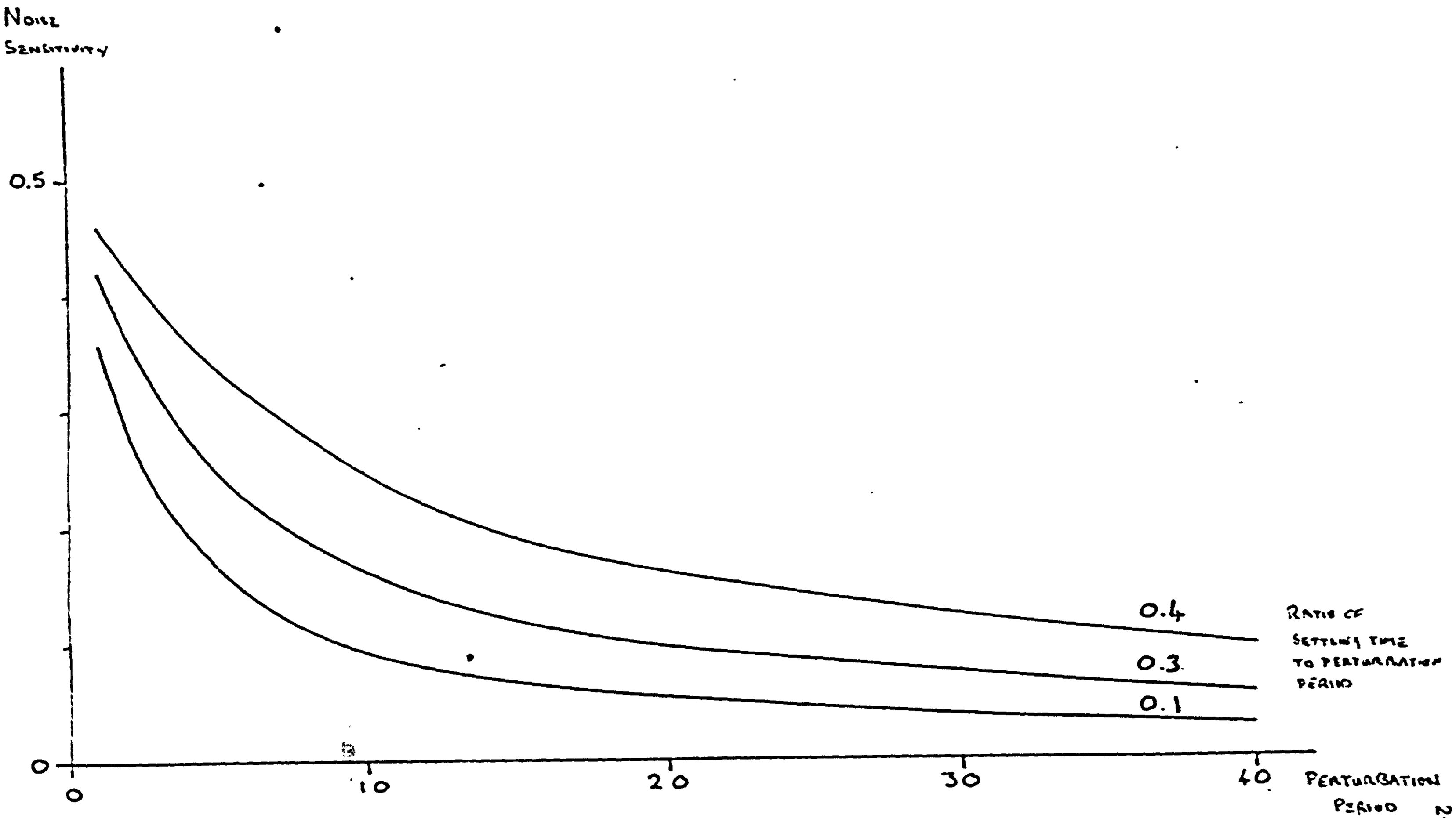


Fig. 2.7.5 - Effect of square wave period on the noise sensitivity of the gain estimator

$$\underline{U_T}'\underline{U_T} = N \begin{bmatrix} \phi_0 & \phi_1 & \phi_2 & \cdot & \cdot & \cdot & \phi_{k-1} \\ \phi_1 & \phi_0 & & & & & \cdot \\ & & \cdot & & & & \\ \phi_2 & & & & & & \cdot \\ & & & \cdot & & & \\ \cdot & & & & & & \cdot \\ & & & & \cdot & & \\ \cdot & & & & & \phi_0 & \phi_1 \\ & & & & & & \\ \phi_{k-1} & \cdot & \cdot & \cdot & \cdot & \phi_1 & \phi_0 \end{bmatrix}$$

where ϕ_m, the autocorrelation function for the sequence $\langle u \rangle$ for shift m, is defined by:

$$\phi_m = -\frac{1}{4N} \sum_{n=1}^{N} \left\{\omega^n - \omega^{-n}\right\}\left\{\omega^{m+n} - \omega^{-(m+n)}\right\}$$

$$= \frac{(\omega^m + \omega^{-m})}{4}$$

The matrix $\underline{U_T}'\underline{U_T}$ is of rank 2 since the sum of the pth and $p + 2$th row equals $(\omega + \omega^{-1})$ times the $P + 1$th row. Solving for $k = 2$,

$$[\underline{U_T}'\underline{U_T}]^{-1} = \frac{-4}{N(\omega - \omega^{-1})^2} \begin{vmatrix} 2 & -(\omega + \omega^{-1}) \\ -(\omega + \omega^{-1}) & 2 \end{vmatrix}$$

Using the result of equations 2.3.4 and 2.3.5,

$$\hat{s} = \frac{2}{N \cos\frac{\pi}{N}} \sum_{n=1}^{N} \sin(2n - 1)\frac{\pi}{N} \cdot y_n$$

and $$\text{var}(\hat{s}) = \frac{2\sigma^2}{N \cos\frac{\pi}{N}}$$

and for unit power,

$$\text{var}(\hat{s}) = \frac{\sigma^2}{N \cos^2\frac{\pi}{N}}$$

which is plotted in fig. 2.7.6 for a range of perturbation periods.

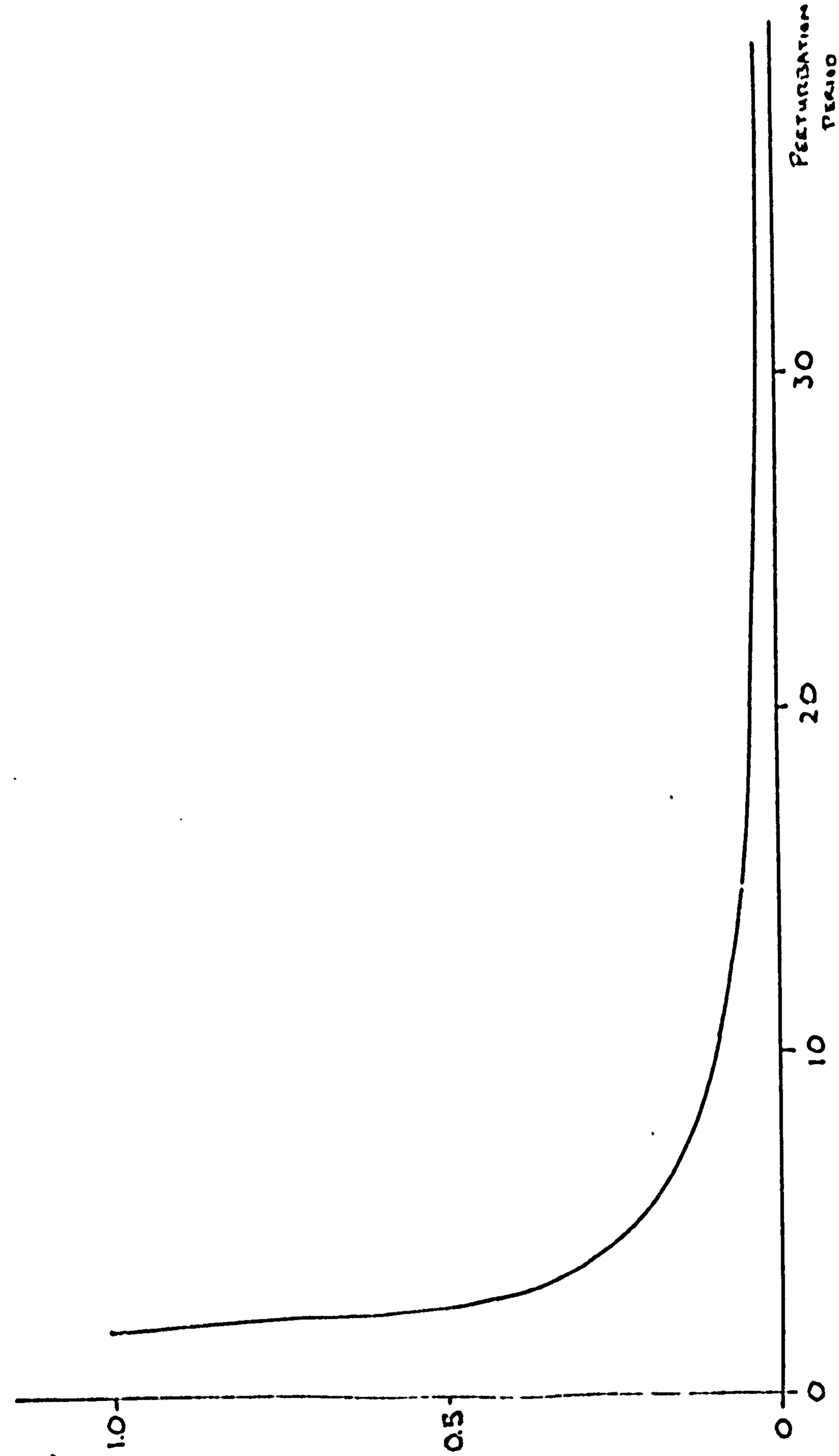


Fig. 2.7.6 – Effect of sine wave period on the noise sensitivity of the gain estimator

2.8 Conclusions

The loop gain of the optimising system has a considerable effect on the dynamic behaviour and the sensitivity to noise. Instability is possible for loop gains less than -2 and for a rapid response the loop gain should be near -1. When noise is present, the lower loop gains give smaller *wander* at the output. The loop gain is given by the product of optimiser gain and the slope of the cost function. In the simple optimiser described, the former is fixed and the latter can normally only be given as an approximate range of values as it is a non-linear function of operating point whose characteristics may change with time. Conservative estimates of the optimiser gain should therefore be made to ensure stability and insensitivity to noise.

The analyses have shown the value of using the matrix method in the general analysis of errors and details of the perturbation waveform, system and noise characteristics allow prediction of optimiser performance.

The choice of perturbation waveform will directly influence the system performance of the presence of noise. Fig. 2.8.1 offers a comparison for some particular waveforms although a direct comparison is difficult as different classes of perturbation exist only at characteristic lengths, which do not necessarily coincide with one another.

Waveforms without rotation symmetry, p.r.b.s. being a specific example, may be used to identify systems with settling times almost

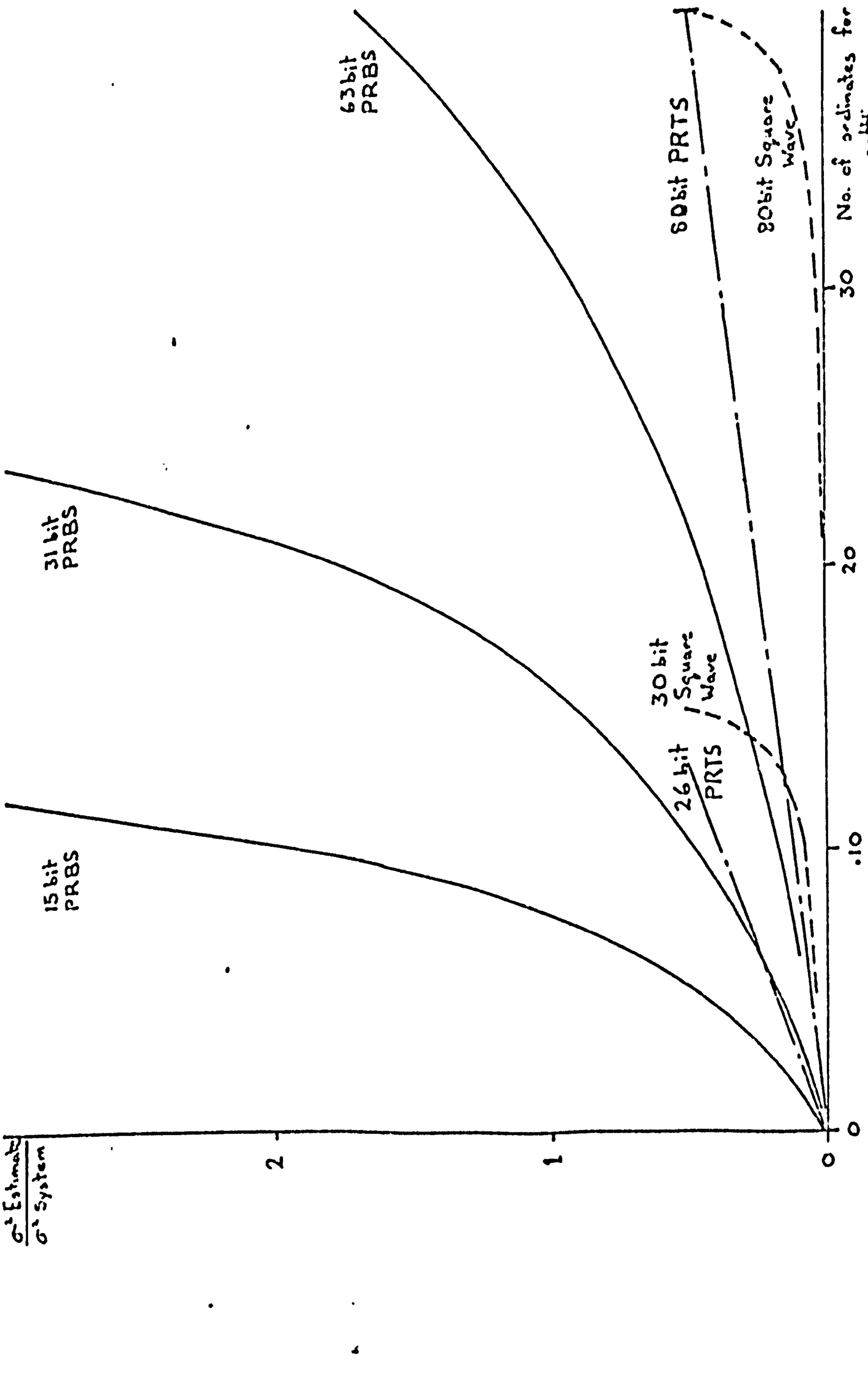


Fig. 2.8.1 - Comparison of noise rejection of gain estimators using a range of perturbations

the length of the sequence. Sequences with rotation symmetry, however, may only be used to identify system settling times up to half the length of the perturbation period and where time is at a premium, this may be a disadvantage.

CHAPTER 3

Further Linear and Non-Linear Optimisers

Introduction

In the simple system of Chapter 2, one value of the gain of the system being optimised, was the only information available, for the estimation of the step to be taken towards the optimum. It was shown that the selection of the optimiser loop gain had a significant effect on the dynamic performance of the system, and that even simple systems with a quadratic cost function could become unstable using this hill climbing method. In this chapter, linear and non-linear methods using more past information about the system are presented for comparison with the simple system. The non-linear methods introduce different estimation problems and these are discussed.

3.1 Higher Order Linear Optimisers

3.1.1 Stability Criteria

Theoretically, the introduction of a zero into the optimiser may result in a stable system for all values of loop gain. It can be seen from the root locus in fig. 3.1.1 that the zero must lie within the unit circle to fulfil this criteria. The transfer function of the optimiser would then have the form,

$$\frac{(z - \beta)}{(z - 1)}, \qquad (3.1.1)$$

but this is physically unrealisable as the numerator is of a higher order in z than the denominator. In general, the system will only be stable for all values of loop gain when particular values for the additional poles and zeros are used and when the number of zeros in the optimiser transfer function exceeds the number of poles by one. This, however, will always give a physically unrealisable form.

The introduction of a further pole into the above transfer function gives the realisable form,

$$\frac{z}{(z - 1)} \cdot \frac{(z - \beta)}{(z - \alpha)}$$

The forward path transfer function for the whole system is then

$$\frac{kg}{(z - 1)} \cdot \frac{(z - \beta)}{(z - \alpha)}$$

and the root locii are shown in fig. 3.1.2, the particular pattern depending on the value of α and β.

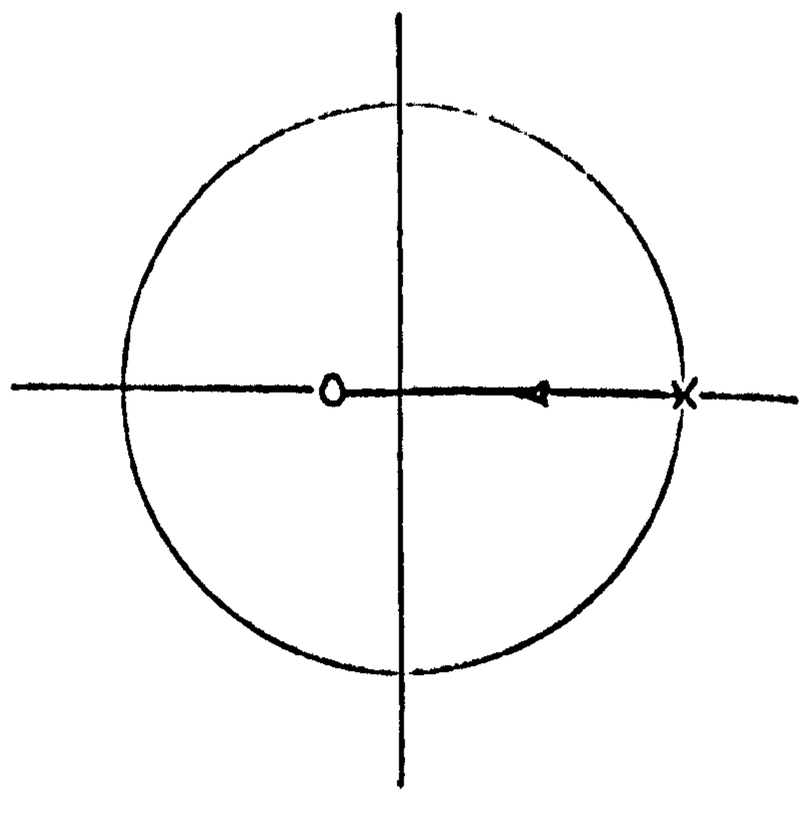


Fig 3.1.1 Root locus with additional zero

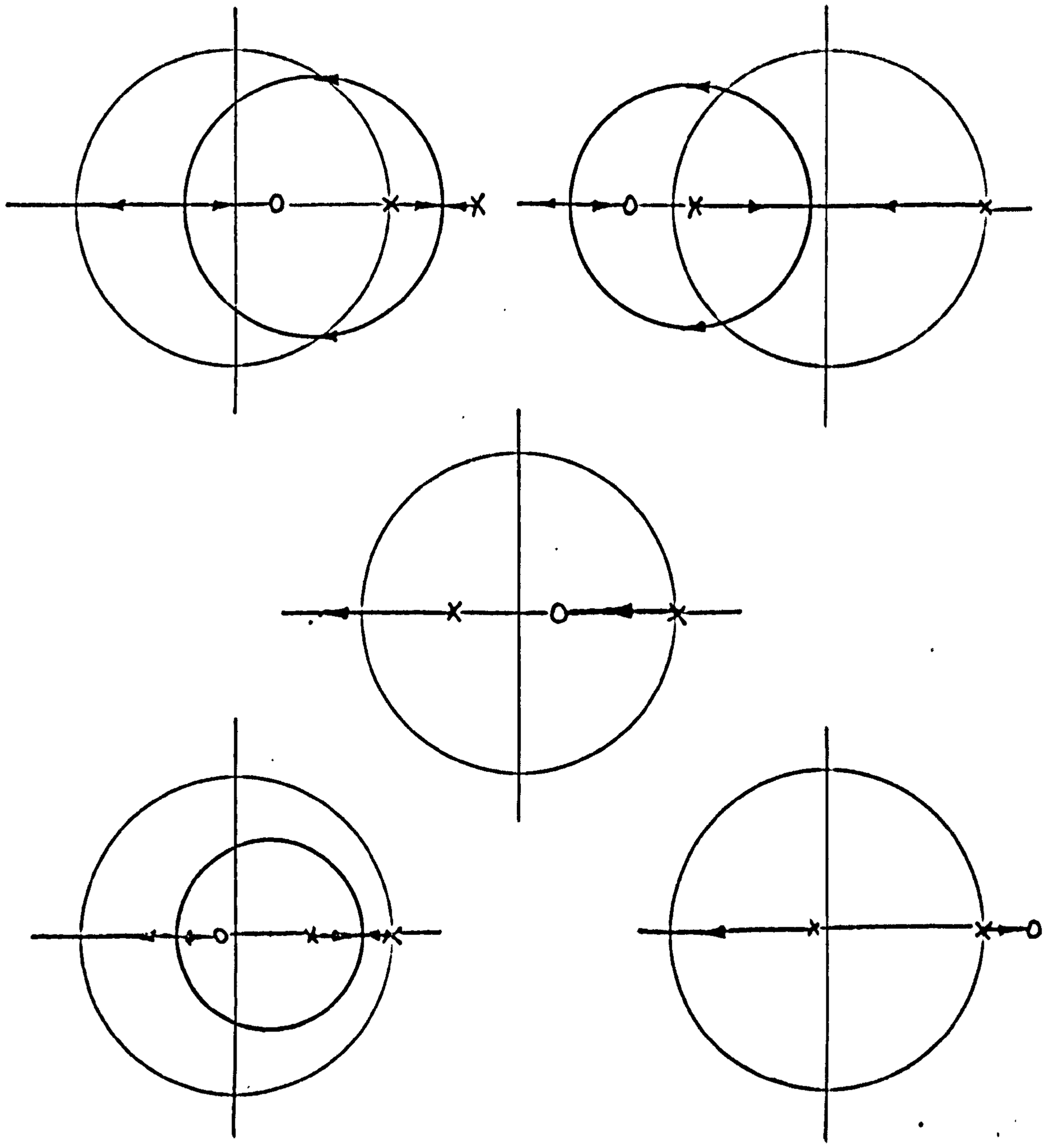


Fig. 3.1.2 Possible root loci of second order linear system

The characteristic equation for the new system is

$$z^2 - (\alpha - 1 - kg)z + \alpha - kg\beta = 0$$

Application of the transform $z = (1 + w) / (1 - w)$ and the Routh-Hurwitz criteria gives the following conditions for stability

$$2(1 + \alpha) - kg(1 + \beta) > 0$$

$$1 - \alpha + kg\beta > 0$$

$$\beta < 1$$

These are illustrated in fig. 3.1.3.

3.1.2 Effect of Noise for a Simplified Case

Consider the simplified case when the zero lies at the origin, $\beta = 0$. The closed loop transfer function then has the form

$$\frac{kgz}{z^2 + (kg - \alpha - 1)z + \alpha}$$

Using the technique of section 2.3.1, the variance of the *wander* of the system due to errors in the gain estimate is

$$\frac{2(1 + \alpha)r^2}{(1 - \alpha)(2(1 + \alpha) - kg)} \qquad (3.1.2)$$

If g is always chosen so that the maximum value of kg gives a stable system with a margin factor γ, the upper limit of gain for a stable system will be $k_{max}g\gamma$. If the current value of k is then a fraction of the maximum k_{max}, qk_{max} where $q \leq 1$, and conditions for stability of this system are

$$\text{loop gain} < 2(1 + \alpha) \text{ and } 1 > \alpha > -1$$

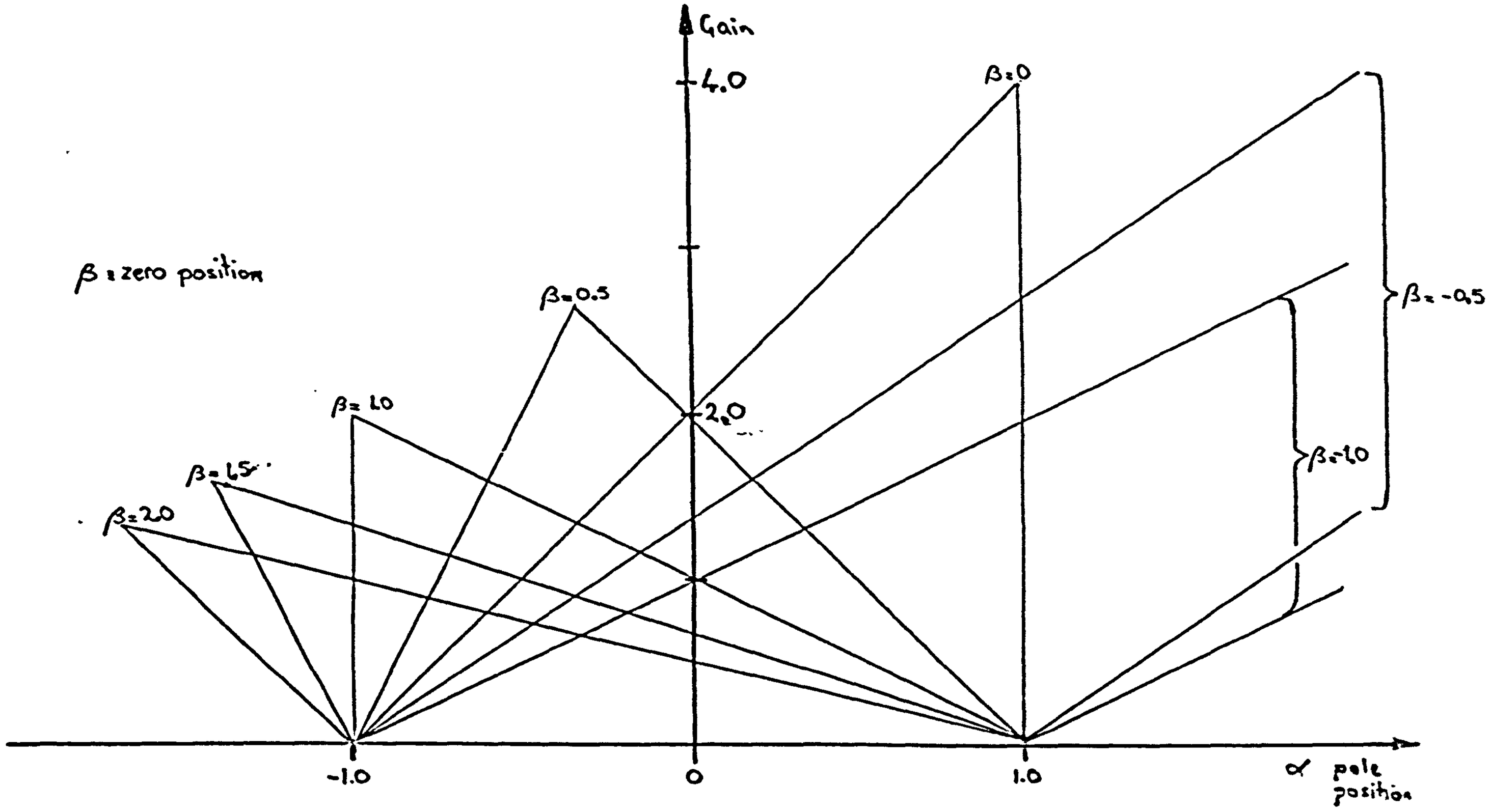


Fig. 3.1.3 Summary of stability criteria. Triangular regions indicate stable operating areas

then

$$k_{max}g\gamma = 2(1 + \alpha)$$

$$g = \frac{2(1 + \alpha)}{\gamma k_{max}}$$

and $$kg = qk_{max}g$$

$$= \frac{2q(1 + \alpha)}{\gamma}$$

Substituting in equation 3.1.2, the variance of the *wander* at the output of the system is given by

$$\frac{1}{(1 - \alpha)} \cdot \frac{\gamma}{(\gamma - q)} \cdot r^2 ,$$

which is plotted in fig. 3.1.4. It can be seen that for small variance, q should be small and γ large, which implies a small value of kg and hence g. In addition, α should be made large and negative but stability criteria will only allow a value greater than -1.

When α is zero, the additional pole cancels with the zero and system reduces to the simple system of Chapter 2.

3.1.3 Dynamic Performance for the Simplified Case

The analysis to determine the dynamic performance of the system has been carried out elsewhere[25] and it has been shown that the actual response is a complex function of the closed loop pole and zero positions.

An approximate guide to the closed loop dynamic performace can however be ascertained by studying the position on the root locus of

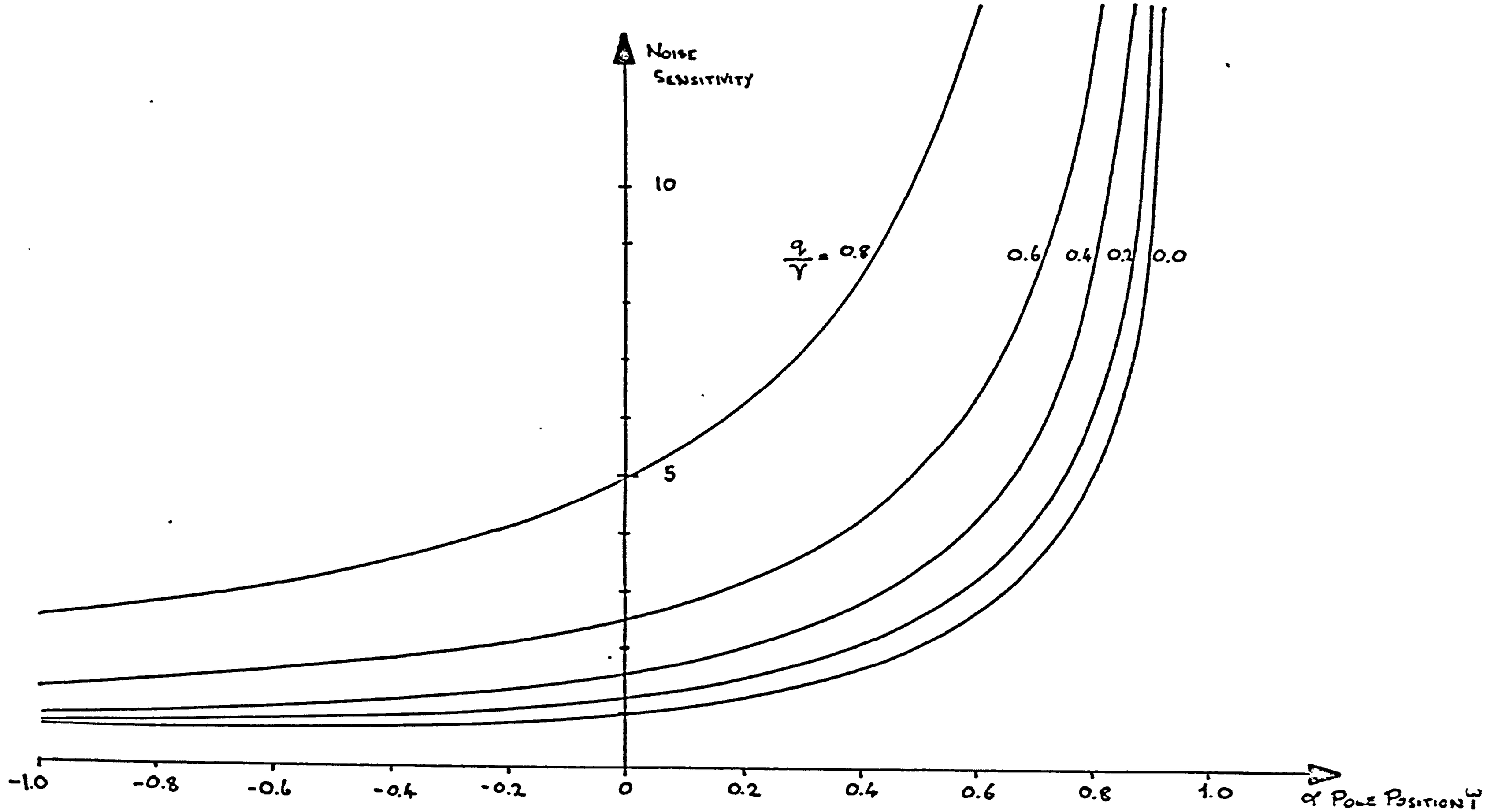


Fig. 3.1.4 Noise sensitivity for second order linear optimiser

the dominant closed loop root or roots of the system. This is directly related to the dominant time constant in the transient response. In the z-plane, the lines of constant time constant, τ, are circles of radius $e^{T/\tau}$, where T is the time between optimiser adjustments. A plot of the radius at which the dominant pole lays versus the proportion of maximum stable gain for a range of values of α is given in fig. 3.1.5 for the system being studied. It can be seen that as α tends to unity, the dominant time constant rapidly decreases as the gain increases, although for values of α greater than zero, a minimum value for the dominant time constant is reached when the system becomes oscillatory.

This suggests that for a short time constant over a wide range of operation, the system should have a value of α between zero and unity. The precise value of α should be chosen so that when the system becomes oscillatory, the time constant is sufficiently short to satisfy the specification of the optimiser.

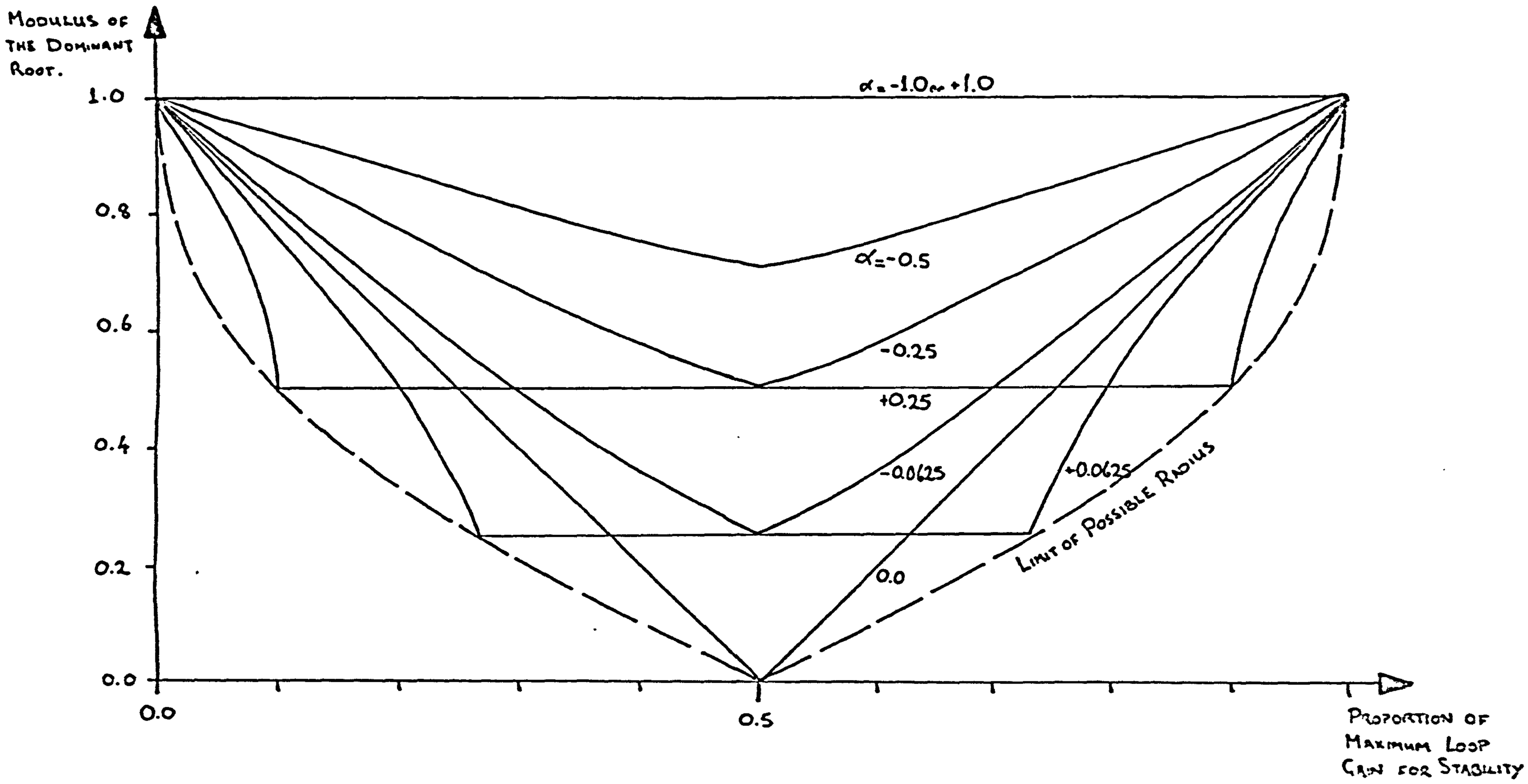


Fig. 3.1.5 Position of dominant root as a function of maximum stable gain

3.2 Non-Linear Optimisers

3.2.1 Application of Newton-Raphson

The optimisation of a system implies searching for an extremum in a cost function. This may be reformulated as locating zero cost function gradient, providing the cost function is continuous. Many of the methods available for locating the zeros of a function require an analytical form but in Newton-Raphson, only the values of the function and its gradient for particular operating points are needed.

Consider a cost function $F(\alpha)$ given by the Taylor expansion,

$$F(\alpha) = F(\alpha_1) + (\alpha - \alpha_1)\left(\frac{\partial F}{\partial \alpha}\right)_{\alpha=\alpha_1} + \frac{(\alpha - \alpha_1)^2}{2!}\left(\frac{\partial^2 F}{\partial \alpha^2}\right)_{\alpha=\alpha_1} + \ldots$$

$$= a_0 + a_1(\alpha - \alpha_1) + a_2(\alpha - \alpha_1)^2 + \ldots \qquad (3.2.1)$$

If a quadratic model is assumed, then the optimum is given by

$$F'(\alpha) = a_1 + 2a_2(\alpha - \alpha_1) = 0$$

and $$(\alpha - \alpha_1) = -\frac{a_1}{2a_2}$$

Thus, if it is possible to determine the first and second derivatives of the cost function at the current operating point α_1, then their ratio gives an estimate of the change in operating point required to reach the optimum. The calculation may then be repeated at the new operating point to give a better estimate of the position of the optimum. It can be seen that for a noise-free system with a quadratic cost function, the optimum will always be located in one step.

3.2.2 Utilisation of 3-level maximal length sequences

It has been shown[26] that it is possible to estimate the first and second derivatives of a hill in one experiment by using a 3-level maximal length sequence perturbation on a particular system configuration. Consider the system shown in fig. 3.2.1 with a perturbation $u(t)$ at an operating point α_1. From equation *3.2.1*, the output of the cost function is given by

$$a_0 + a_1 u(t) + a_2 u^2(t)$$

The linear and square law terms of the non-linearity may be separated to give a system with two parallel paths, so that when the system is perturbed by a 3-level sequence, the linear dynamics of the first path are perturbed by a proportion of the sequence and the dynamics of the second path by a different proportion of the square of the sequence.

Let the weighting sequence for the linear dynamics at the operating point including the first derivative scaling factor a_1 be $\underline{h}$ and that of the equivalent second derivative path be $\underline{g}$ where

$$\underline{g} = \frac{a_2}{a_1} \underline{h} ,$$

a_2 being the magnitude of the second derivative at the operating point. The output of the system due to the perturbation is then given by

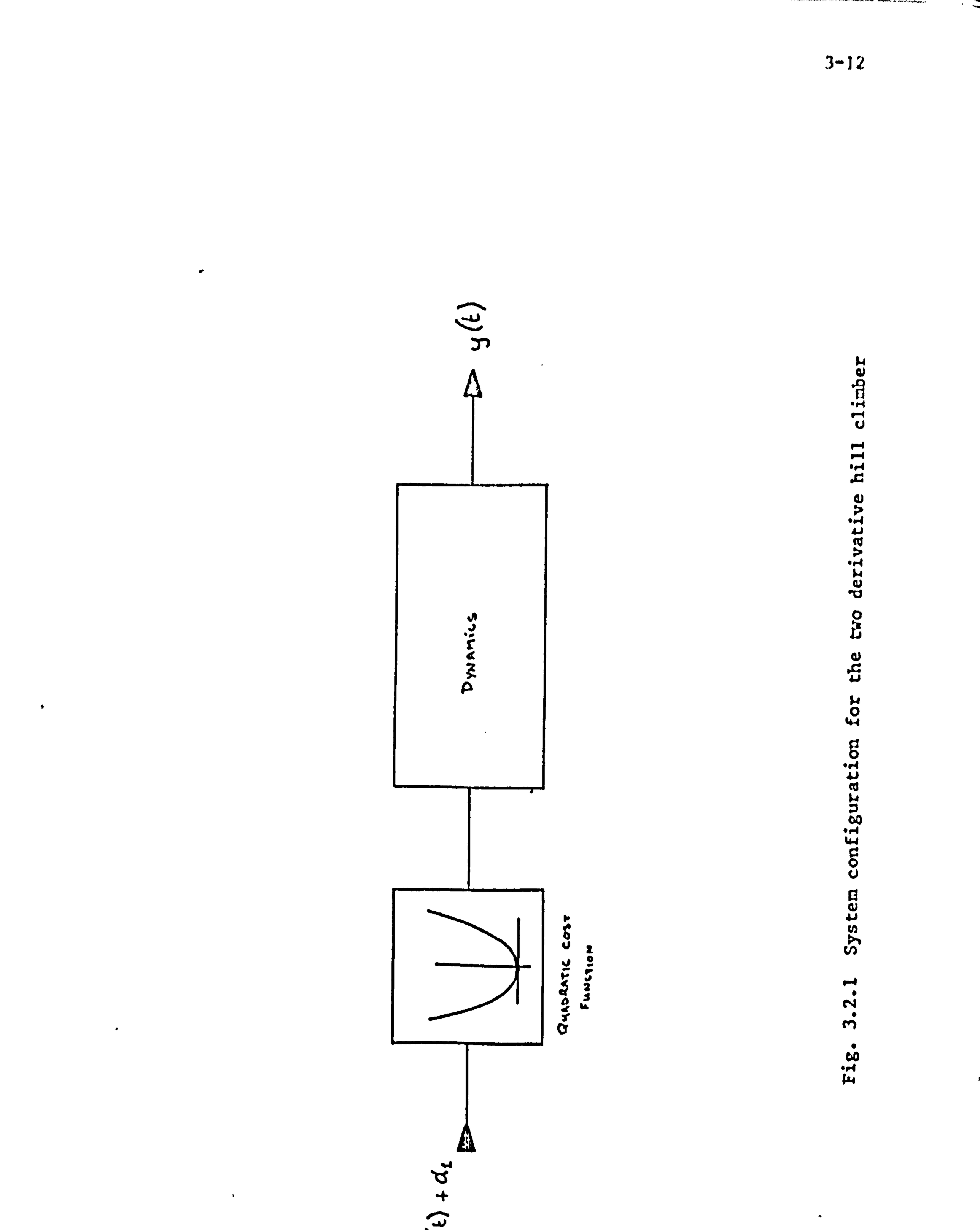


Fig. 3.2.1 System configuration for the two derivative hill climbe

$$\underline{u} = \begin{bmatrix} u_N^2 & . & . & . & u_1^2 \\ u_1^2 & u_N^2 & . & . & u_2^2 \\ . & & . & & . \\ . & & . & & . \\ u_{N-1}^2 & & & . & u_N^2 \end{bmatrix} \begin{bmatrix} g_1 \\ . \\ . \\ . \\ g_N \end{bmatrix} + \begin{bmatrix} u_N & . & . & . & u_1 \\ u_1 & u_N & & & u_2 \\ . & & & & \\ . & & & & \\ u_{N-1} & & & & u_N \end{bmatrix} \begin{bmatrix} h_1 \\ . \\ . \\ . \\ h_N \end{bmatrix}$$

(3.2.2)

It is not possible to estimate $\underline{g}$ and $\underline{h}$ from this equation set as there are $2N$ unknowns and only N equations. If, however, it is assumed that the system settles in $N/2$ intervals, then the last $N/2$ values of $\underline{g}$ and $\underline{h}$ will be zero and equation 3.2.2 may be written as

$$\underline{y} = \left[\begin{array}{cccc|cccc} u_N^2 & u_{N-1}^2 & . & u_{N/2}^2 & u_N & u_{N-1} & . & u_{N/2} \\ u_1^2 & u_N^2 & & . & u_1 & u_N & & . \\ . & & & . & . & & & . \\ . & & & . & . & & & . \\ . & & & . & . & & & . \\ u_{N-1} & & & u_N^2 & u_{N-1} & & & u_N \end{array}\right] \left[\begin{array}{c} g_1 \\ . \\ . \\ \underline{g_{N/2}} \\ h_1 \\ . \\ . \\ . \\ h_{N/2} \end{array}\right]$$

$$= \underline{P} \begin{bmatrix} \underline{g} \\ \hline \underline{h} \end{bmatrix} \text{ say}$$

This may now be solved to give

$$\begin{bmatrix} \hat{\underline{g}} \\ \hline \hat{\underline{h}} \end{bmatrix} = [\underline{P}'\underline{P}]^{-1}\underline{P}'\underline{y} \qquad (3.2.3)$$

3.2.3 Introduction of Steady State System Output

If the operating point and inherent system bias terms are included, the estimate of the first derivative path dynamics will be unaltered but the second derivative path estimate will be shifted by a constant. Then

$$\hat{\underline{g}} = \underline{g} + \text{constant}$$

$$\hat{\underline{h}} = \underline{h}$$

and $$\hat{\underline{h}} = \frac{a_1}{a_2}\hat{\underline{g}} + \gamma 1 \ , \qquad (3.2.4)$$

where γ is a constant. Equation *3.2.4* is a straight line with slope r given by the ratio of the first derivative to the second.

A simulation was carried out on an analogue computer for a second order system preceded by a quadratic non-linearity. The estimates of the first and second derivative path dynamics are shown in fig. 3.2.2, the effect of the steady state level having been removed from the second derivative. Fig. 3.2.3 shows these estimates plotted against one another with time as a parameter for different ratios of the first to the second derivative.

3.2.4 Effects of System Noise

If noise is also present at the system output, then the estimates of $\underline{g}$ and $\underline{h}$ will be contaminated by noise $\underline{\xi}_g$ and $\underline{\xi}_h$ respectively and the estimated weighting sequences will be

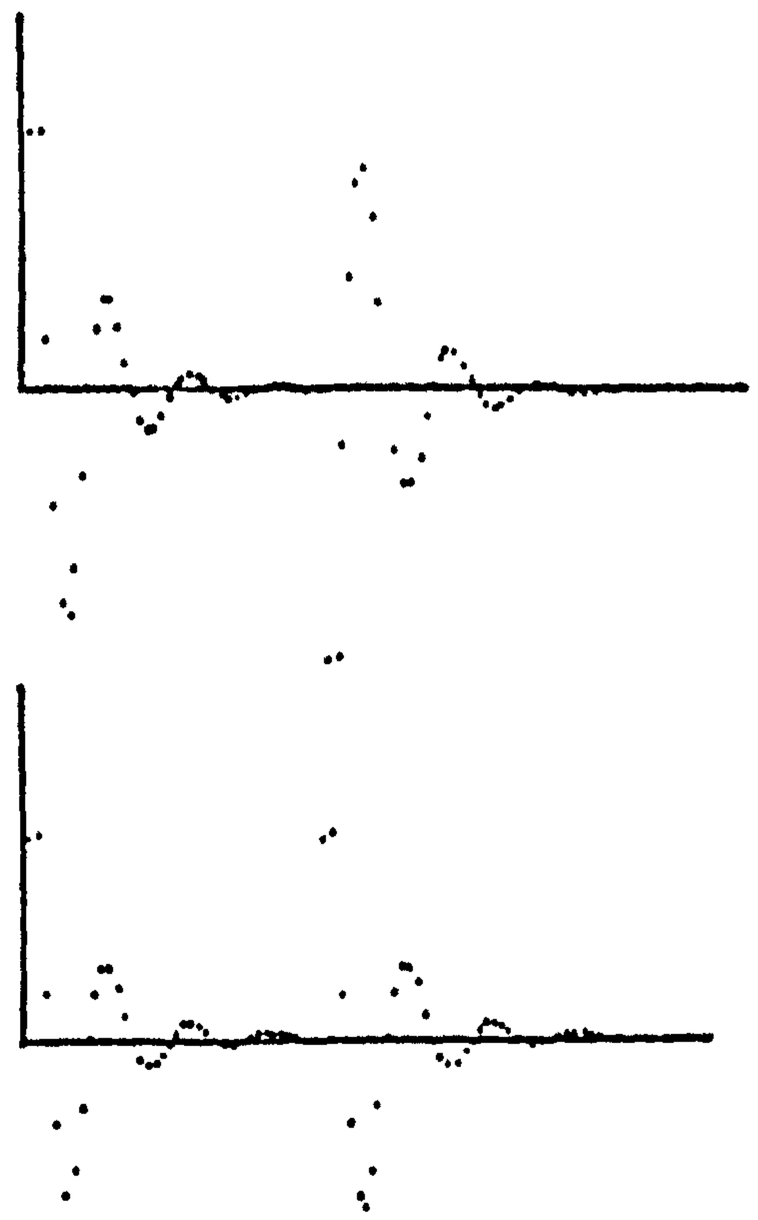


FIG 3.2.2. ESTIMATES OF (a) FIRST DERIVATIVE PATH DYNAMICS
(b) SECOND DERIVATIVE PATH DYNAMICS
FOR A SIMULATED NON-LINEAR SECOND ORDER SYSTEM.

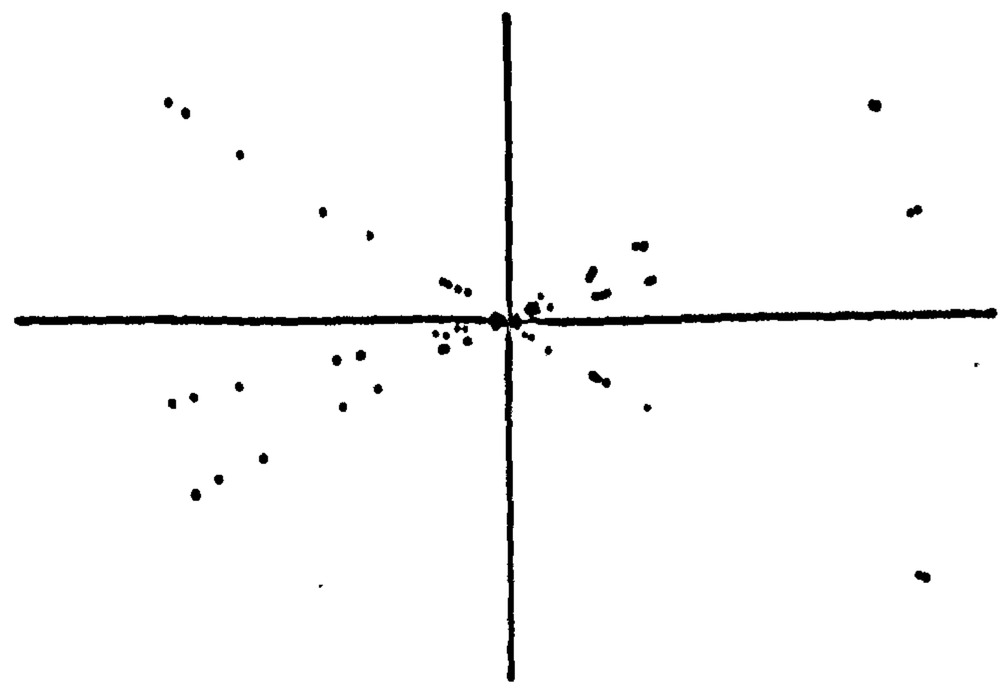


FIG 3.2.3. FIRST AND SECOND DERIVATIVE PATH IMPULSE RESPONSE ESTIMATES PLOTTED AGAINST ONE ANOTHER FOR THREE FIRST DERIVATIVE VALUES.

$$\hat{\underline{g}} = \underline{g} + \underline{\xi}_g + \text{constant} \qquad (3.2.5)$$

$$\underline{\hat{h}} = \underline{h} + \underline{\xi}_h \qquad (3.2.6)$$

Using least squares, the estimated value for $\frac{a_1}{a_2}$ is given by

$$\hat{r} = \underline{\hat{h}}' \left[\underline{I} - \frac{2}{N}\underline{J}\right]\hat{\underline{g}}\left[\hat{\underline{g}}' \left[\underline{I} - \frac{2}{N}\underline{J}\right]\hat{\underline{g}}\right]^{-1} \qquad (3.2.7)$$

From equations 3.2.4, 3.2.5 and 3.2.6,

$$\underline{\hat{h}} = r\hat{\underline{g}} + \gamma\underline{1} + \underline{\xi}_h - r\underline{\xi}_g$$

and the covariance of $\underline{g}$ and the errors is given by

$$E\left\{\left[\hat{\underline{g}} - E(\hat{\underline{g}})\right]' \left[\underline{\xi}_h - r\underline{\xi}_g\right]\right\}$$

$$= E\left\{\underline{\xi}_g' \left[\underline{\xi}_h - r\underline{\xi}_g\right]\right\}$$

$$= -r \text{ covar } (\underline{\xi}_g)$$

This implies that the values of $\hat{\underline{g}}$ are correlated with the errors and therefore the application of least squares will lead to a biased estimate for the value of r.

Substituting from equations 3.2.5 and 3.2.6 in equation 3.2.7 and simply assuming the errors $\underline{\xi}_g$ and $\underline{\xi}_h$ are independent of each other and the true values of $\underline{g}$ and $\underline{h}$,

$$\hat{r} = \frac{\underline{g}' \left[\underline{I} - \frac{2}{N}\underline{J}\right]\underline{h}}{\underline{g}' \left[\underline{I} - \frac{2}{N}\underline{J}\right]\underline{g} + \xi_g' \left[\underline{I} - \frac{2}{N}\underline{J}\right]\xi_g}$$

For a large sample

$$\underset{N \to \infty}{\text{plim}}\ \hat{r} = \frac{r}{1 + K},$$

where $K = \underline{\xi}_g' \left[\underline{I} - \frac{2}{N}\underline{J}\right]\underline{\xi}_g \left[\underline{g}' \left[\underline{I} - \frac{2}{N}\underline{J}\right]\underline{g}\right]^{-1}$

There-fore for large samples, the estimate of r is a function of the variance of the values of g and the variance of the noise in g and the probability limit will not converge on r. The least squares estimate of r is thus inconsistent.

3.2.5 Special Case of Least Squares

If the line is constrained to pass through the origin, then the least squares estimate of r reduces to

$$\hat{r} = \underline{1}'\underline{h}[\underline{1}'\underline{\hat{g}}]^{-1} \qquad (3.2.8)$$

This estimate will still be biased as

$$\underline{h} = r\underline{g}$$

Expanding equation 3.2.8 however,

$$\hat{r} = \underline{1}'[\underline{h} + \underline{\xi}_h][\underline{1}'[\underline{g} + \underline{\xi}_g]]^{-1}$$

$$\underline{1}'[r\underline{g} + \underline{\xi}_h][\underline{1}'[\underline{g} + \underline{\xi}_g]]^{-1}$$

In this case, for a large sample

$$\underset{N \to \infty}{\text{plim}}\ \hat{r} = r,$$

providing $E\{\underline{\xi}_g\} = E\{\underline{\xi}_h\} = 0$.

Therefore when the line is constrained to pass through the origin, the least squares estimate of r is biased but consistent.

This method of estimation may be implemented if the system can be assumed to have settled in less than half the perturbation period. The later ordinates may then be used to remove the effect of steady state bias on the second derivative estimate. A disadvantage of this procedure is that the perturbation length must be greater than twice the settling time of the system. Godfrey and Clarke have utilized this method of estimation in a hill climber[27].

3.2.6 Maximum Likelihood Estimate

An alternative approach is to use a maximum likelihood method which ensures that the estimate is always consistent although it may still be biased. Johnston[5] gives the solution for sample errors with variance only, but for a general covariance matrix the likelihood function is given by

$$L \propto \det(\sigma^2 \underline{G})^{-N/4} \exp\left\{-\tfrac{1}{2}[\hat{\underline{q}} - \underline{q}]'[\sigma^2 \underline{G}]^{-1}[\hat{\underline{q}} - \underline{q}]\right\}$$

$$\det(\sigma^2 \underline{H})^{-N/4} \exp\left\{-\tfrac{1}{2}[\hat{\underline{h}} - \alpha \underline{1} - r\underline{q}]'[\sigma^2 \underline{H}]^{-1}[\hat{\underline{h}} - \alpha \underline{1} - r\underline{q}]\right\},$$

where $\sigma^2 \underline{G}$ and $\sigma^2 \underline{H}$ are the covariance matrices for $\hat{\underline{q}}$ and $\hat{\underline{h}}$ respectively. The log likelihood function L^* is then given by

$$L^* = \text{constant} - \frac{N}{2} \log\sigma^2 - \frac{1}{2\sigma^2}[\hat{\underline{q}} - \underline{q}][\sigma^2 \underline{G}]^{-1}[\hat{\underline{q}} - \underline{q}]$$

$$- \frac{1}{2\sigma^2}[\hat{\underline{h}} - \alpha \underline{1} - r\underline{q}]'[\sigma^2 \underline{H}]^{-1}[\hat{\underline{h}} - \alpha \underline{1} - r\underline{q}]$$

Taking partial derivatives with respect to α, $\underline{g}$ and r and equating to zero to give the maximum likelihood estimates $\tilde{\alpha}$, $\underline{\tilde{g}}$ and $\tilde{r}$,

$$\left.\begin{aligned} \underline{1}'\underline{H}^{-1}[\underline{\hat{h}} - \tilde{r}\underline{\tilde{g}} - \tilde{\alpha}\underline{1}] &= 0 \\ \underline{\tilde{g}}'H^{-1}[\hat{h} - \tilde{r}\underline{\tilde{g}} - \tilde{\alpha}\underline{1}] &= 0 \\ G^{-1}[\hat{\underline{g}} - \underline{\tilde{g}}] + \tilde{r}\underline{H}^{-1}[\underline{\hat{h}} - \tilde{r}\underline{\tilde{g}} - \tilde{\alpha}\underline{1}] &= 0 \end{aligned}\right\} \qquad (3.2.9)$$

For a 3-level maximal length sequence, $\underline{P}'\underline{P}$ in equation 3.2.3 is given by

$$\underline{P}'\underline{P} = \left[\begin{array}{c|c} \frac{1}{3}[\underline{I} + 2\underline{J}] & \underline{0} \\ \hline \underline{0} & \underline{I} \end{array}\right]$$

and

$$[\underline{P}'\underline{P}]^{-1} = \left[\begin{array}{c|c} \underline{G} & \underline{0} \\ \hline \underline{0} & \underline{H} \end{array}\right]$$

$$= \left[\begin{array}{c|c} 3\left[\underline{I} - \frac{2}{(N+1)}\underline{J}\right] & \underline{0} \\ \hline \underline{0} & \underline{I} \end{array}\right]$$

Substituting for $\underline{G}^{-1}$ and $\underline{H}^{-1}$ in the equation set 3.2.9 and simplifying.

$$\underline{\tilde{g}} = \left[\tfrac{1}{3}[\underline{I} + 2\underline{J}] + \tilde{r}^2\left[\underline{I} - \tfrac{2}{N}\underline{J}\right]\right]^{-1}\left[\tfrac{1}{3}[\underline{I} + 2\underline{J}]\underline{\hat{g}} + \tilde{r}\left[\underline{I} - \tfrac{2}{N}\underline{J}\right]\underline{\hat{h}}\right]$$

$$\underline{\tilde{g}}'\left[\underline{I} - \tfrac{2}{N}\underline{J}\right]\underline{\hat{h}} - \tilde{r}\underline{\tilde{g}}'\left[\underline{I} - \tfrac{2}{N}\underline{J}\right]\underline{\tilde{g}} = 0$$

Combining these two equations gives

$$3\tilde{r}^2 M_{gh} + \tilde{r}(M_{gg} - 3M_{hh}) - M_{gh} = 0 \qquad (3.2.10)$$

where $M_{gh} = \underline{\hat{g}}'\left[\underline{I} - \tfrac{2}{N}\underline{J}\right]\underline{\hat{h}}$

$$M_{gg} = \underline{\hat{g}}'\left[\underline{I} - \tfrac{2}{N}\underline{J}\right]\underline{\hat{g}}$$

and $$M_{hh} = \underline{\hat{h}}'\left[\underline{I} - \tfrac{2}{N}\underline{J}\right]\underline{\hat{h}}$$

Equation *3.2.10* may be solved to give the maximum likelihood estimate of r.

By substituting from equation 3.2.3, it is possible to determine the quantities M_{gh}, M_{gg} and M_{hh} directly from the system response without the intermediate computation of $\hat{g}$ and $\underline{h}$.

3.2.7 Effect of an alternative system configuration

Consider the configuration shown in fig. 3.2.4 where the cost function is assumed to be separable from the dynamics. The response to the perturbation due to the square law in the cost function is then given by

$$y(t) = a_2 \int_0^\infty h_2(x_3)\Big\{ \int_0^\infty h_1(x_1)u(t - x_3 - x_1)dx_1 \int_0^\infty h_1(x_2)u(t - x_2 - x_1)dx_2\Big\}dx_3$$

Correlating with the square of the perturbation,

$$\int_0^T y(t)u^2(t - \tau)dt$$

$$= a_2\int_0^\infty h_2(x_3)\int_0^\infty h_1(x_1) \int_0^\infty h_1(x_2)$$

$$\Big\{\int_0^T u^2(t - \tau)x(t - x_1)u(t - x_2)dt\Big\}dx_2dx_1dx_3$$

where T is the perturbation period.

The fourth order auto-correlation of a three level sequence is dependent on the generating polynomial and is a complicated function

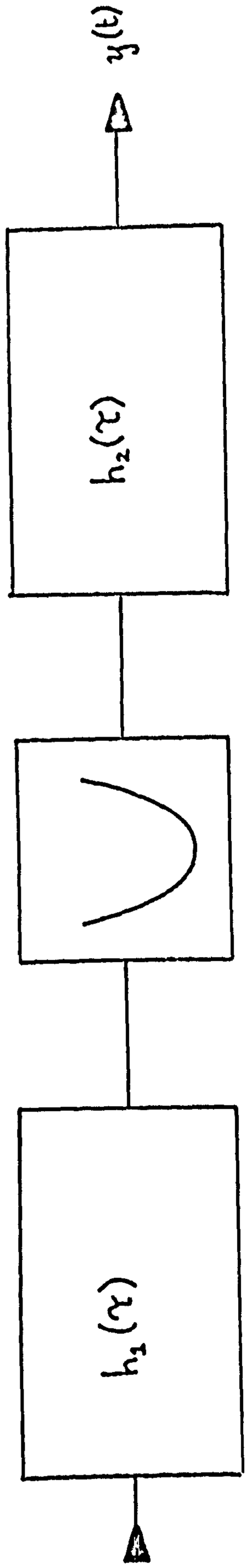


Fig. 3.2.4 Alternative system configuration

of the shifts x_1 and x_2 (Appendix A3). It is therefore impractical to use this expression to compute the magnitude of the second derivative of the cost function.

Fig. 3.2.5 shows the effect of the position of the cost function for a system whose dynamic components were two first order lags with time constants τ_1 and τ_2.

3.2.8 Difference Methods

If the gain estimates G_1 and G_2 at operating points α_1 and α_2 are available, then an estimate of the second derivative is given by

$$\frac{G_1 - G_2}{(\alpha_1 - \alpha_2)}$$

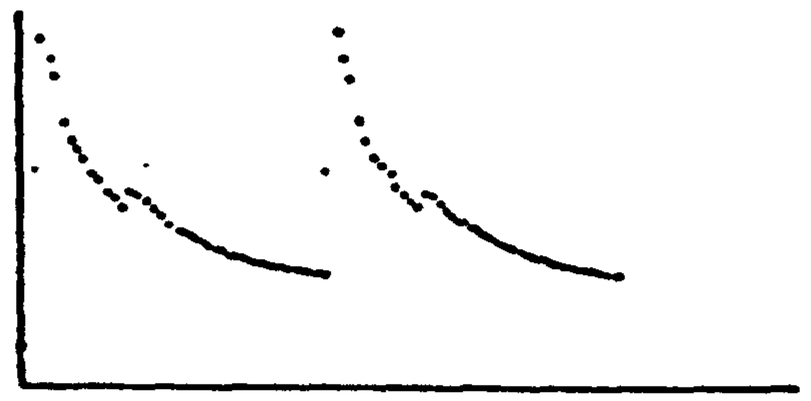
and the modified Newton-Raphson gives the estimated change in operating point as

$$- \frac{(\alpha_2 - \alpha_1)}{1 - G_1 / G_2} \qquad (3.2.11)$$

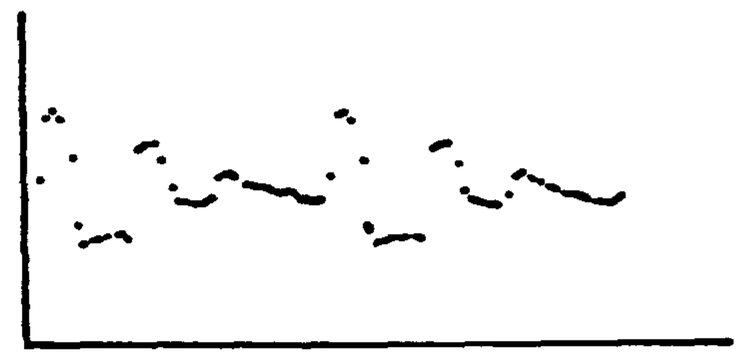
The estimate of the ratio of G_1 to G_2, R, may be made independent of any bias by plotting the two impulse responses against one another and then using a maximum likelihood estimator. Irrespective of the perturbation used, this reduces to a comparison of the outputs of the system at the two operating points. Then the estimate of R is given by the quadratic

$$\tilde{R}^2 M_{P_1P_2} + \tilde{R}\left(M_{P_1P_1} - M_{P_2P_2}\right) - M_{P_1P_2} = 0 \ ,$$

where $\underline{P}_1$ is the output sequence at α_1, $\underline{P}_2$ is the output at α_2 and

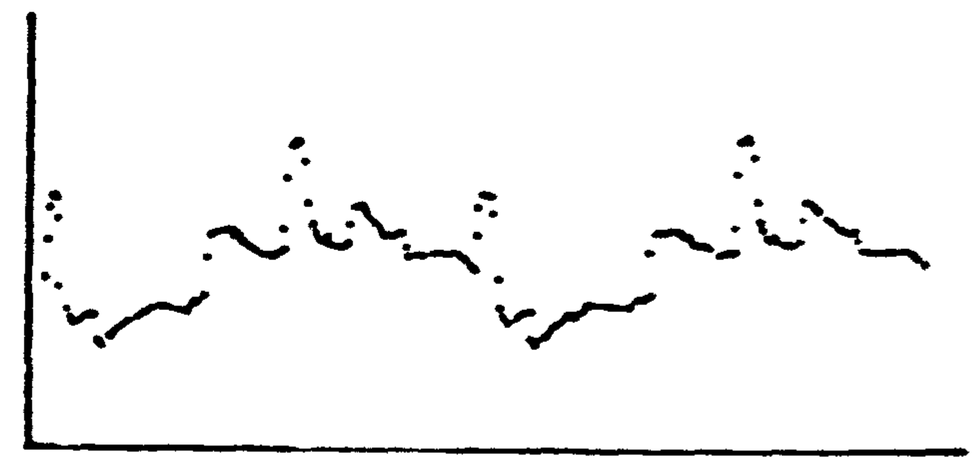


(a)



(b)

(c)



(d)

FIG 3.2.5. EFFECT OF NONLINEARITY ON SECOND DERIVATIVE PATH DYNAMICS ESTIMATES
(a) and (b) USE AN 80 BIT PRTS (c) and (d) A 242 BIT SEQUENCE
(a) and (c) $\tau_1 = 0.01\,\tau_2$ (b) and (d) $\tau_1 = 0.1\,\tau_2$
$\tau_2 \equiv$ 10 BITS OF THE SEQUENCE

$$M_{P_1P_2} = \hat{P}_1 \left[\underline{I} - \frac{2}{N}\underline{J}\right]\hat{P}_2$$

$$M_{P_1P_1} = \hat{P}_1 \left[\underline{I} - \frac{2}{N}\underline{J}\right]\hat{P}_1$$

and $$M_{P_2P_2} = \hat{P}_2 \left[\underline{I} - \frac{2}{N}\underline{J}\right]\hat{P}_2$$

3.2.9 Higher Order Cost Function Models

As the Newton-Raphson hill climbing method is based on a quadratic model of the cost function, a higher order model should give a better performance when the cost function is of a higher order. Such a model may be constructed from estimated values of the cost function; its first and second derivatives at a range of operating points by using difference methods or least squares to fit a higher order polynomial. The latter may only be applied when there is an excess of information. The optimum may then be found analytically or by locating the zeros of the cost function gradient iteratively. Since a higher order model implies more extremums, care must be taken as noisy sytems may introduce additional false optimums near the true optimum.

In general, a higher order model will require more past information, but past values may have to be discarded if the hill shifts or the order of the model is insufficient at the new operating point. This may be carried out systematically by using weighted least squares to give the more recent values greater weight. The weights chosen will be arbitrary, however, unless the movement of the hill or the insufficiency of the model are well known.

3.3 Conclusions

Both the linear and non-linear optimisers discussed in this chapter introduce additional computation when compared with the basic optimiser of Chapter 2. The linear optimisers offer no great advantage as the requirements of noise rejection and speed of response produce opposing criteria for the optimiser parameters, and the simple system lies between these two extremes. The non-linear optimisers with a higher order cost function model require considerable extra computation and introduce many practical difficulties. The two-derivative hill climber however represents a simple method for obtaining a stable system over a wide range of cost functions. The estimation problems for three-level maximal-length sequences have been solved but their use is severely restricted to particular system configurations.

CHAPTER 4

Engine Instrumentation, Modelling and Programmes

Introduction

Practical tests on a working process were carried out to verify the applicability of the theoretical results previously obtained. An engine test rig proved to be the most suitable process available, and previous studies by Draper and Li[3] had already shown that valuable results could be obtained by studying optimisation schemes on an internal combustion engine.

The engine was a 1725 c.c. petrol engine representative of the type currently in use to power medium sized passenger cars, with a maximum power output of 80 b.h.p. occurring at 5000 r.p.m.. It had a conventional four cylinder in-line layout with overhead valves operated by push rods and an aluminium cylinder head. A shaft and two rubber couplings connected the engine to an eddy current dynamometer whose casing was free to swing in trunion bearings and was restrained by a spring balance, to give an indication of the transmitted torque. Because of its mode of operation, the dynamometer was not capable of supplying power to the engine under steady state conditions. A constant stream of water supplied under pressure from a pump removed the heat generated in the dynamometer stator when the dynamometer absorbed power. The engine test cell was equipped with a high pressure cooling water main connected to heat exchangers to remove heat from the dynamometer and engine cooling water circuits and the lubricating oil cooler. Safety circuits shut down the plant

and turned off the engine ignition in the event of a dynamometer cooling water or lubrication failure, or a mains power failure.

A small process control computer was available to run the experiments. The machine had an 8K core store with a twelve bit word length and a cycle time of 1·75μs. Special features included a hardware multiplier and divider, a priority interrupt system with twelve levels of interrupt, an analogue input converter which could be switched to twenty four channels through a hundred points per second random access multiplexer, twelve relays, six analogue output channels and eight single bit digital input channels. As no suitable software was available, programmes were written for the on-line real-time operation of the test rig, using a symbolic assembly language.

Because the engine was insufficiently instrumented to carry out the experimental work, further instruments were designed and constructed, and the installed instruments modified to comply with the requirements of the computer interface. A dynamic model of the rig was developed to assist in the design of the controllers and optimisers.

4.1 Engine Instrumentation

4.1.1

It was essential to control the throttle and load for engine operation, and the ignition angle and fuel-air ratio for the optimisation studies. Actuators for the load, ignition angle and fuel-air ratio had to be developed but the carburettor was already equipped with a feedback angular position controller to operate the throttle. A directly connected stepping motor rotated the throttle and a potentiometer measured its position; when the error between the demanded and measured throttle angle exceeded one step of the driving motor, the motor fields commutated in the correct sense; the maximum commutation rate corresponding to the speed at which the rotor and its load could stop within one step. For small amplitude sinusoidal inputs, up to 5% of full scale, the servo had a flat response up to 15 Hz.

No other variables could be controlled as the computer only had six digital to analogue converters and the remaining two were required by the X-Y plotter. Because of this restriction, the existing oil and water temperature analogue control loops were retained and this also permitted the rig to be operated without the computer. A heat-exchanger with a by-pass loop replaced the vehicle radiator, the engine-driven water pump circulated the cooling water and a thermistor probe measured the water temperature at the engine outlet. A motorised mixing valve controlled the proportion of the flow passing through the heat-exchanger and a feedback controller operating on temperature error adjusted the valve position.

The test engine had provision for an oil cooler to be fitted between the oil pump and the feed to the main bearings. A heat exchanger was installed and the measurement and control of oil temperature carried out in the same way as for the cooling water circuit.

Measurements of torque and speed were also necessary. The torque could be measured from a balance attached to the dynamometer casing, but this had a poor dynamic response. An improved torque measuring instrumentwas therefore developed and the dynamometer casing clamped in subsequent experiments. An a.c. tachometer provided a voltage proportional to the speed which had to be rectified, smoothed and appropriately scaled, using an operational amplifier, to give a signal compatible with the computer interface.

Additional switching and interlocks were introduced to allow the complete remote operation of the engine.

4.1.2 Dynamometer Field Drive

The dynamometer which provided a load for the engine was an eddy current brake[22, 23] of the inductor type[24]. The axially dentated rotor was connected to the engine and a fixed coil, concentric with the machine, supplied an axial field. The changes in permeance, which occurred between a point on the inside of the stator and the rotor when the rotor was in motion, caused an alternating flux to be superimposed on the steady state flux at the stator surface. Additional current in the concentric field coil resulted in a larger ambient flux, which caused an increase in the alternating component of the flux on the stator surface and consequently in the

eddy current losses in the stator. The increase in circulating currents created an increased braking torque on the engine. The drive amplifier for the highly inductive field coil employed a high voltage supply and current feedback to achieve rapid changes in coil current. Small signal step changes in the demanded field current resulted in actual changes in current having a rise time of 4 ms in the completed device.

4.1.3 Electronic Ignition Timing

Control of the ignition timing by electro-mechanical or pneumatic devices which rotate the contact-breaker points was considered inadequate due to the *wander* inherent in mechanical contact-breaker systems. Replacing the contact breaker by reed relays and an electronic ignition circuit gave little improvement over the conventional system. Consequently electronic methods of control were examined and the problem was resolved into finding a method of accurately generating a delay corresponding to a known angle of crankshaft rotation. This may be achieved by analogue or digital methods, the latter being chosen since it was not subject to drift and would operate from cranking speeds up to the maximum engine speed. The installed system provided an efficient means of ignition timing control over a 64° range in $\frac{1}{4}^{\circ}$ steps, and the mode of operation permitted full scale alterations of timing between adjacent firings. A more detailed description of this system is given in Appendix A2.1.

4.1.4 Ignition Firing Unit

The ignition unit had to be replaced by one which would accept a voltage pulse from the ignition timing device. Conventional ignition systems obtain the energy for firing the spark from the current flowing in the primary inductor of the ignition coil. The mechanical contact breaker achieves rapid circuit-breaking with a very high on-off resistance ratio ensuring that the energy is rapidly and almost entirely dissipated in the spark.

Semi-conductor devices cannot attain the required switching performance at the low battery voltage. A capacitor charged to a high voltage was used therefore as an energy store, the capacitor value and voltage being selected to give the same quantity of energy as used in the conventional system. The energy from the capacitor was dissipated in a spark using the ignition coil as a transformer and a thyristor as a switch. The thyristor was triggered at the required instant either by the conventional contact breaker or by a pulse provided electronically. A current limit in the supply restricted the firing rate and prevented overspeeding of the engine.

4.1.5 Fuel Air Ratio

The main jet was accessible through the base of the carburettor and could be raised or lowered by screwing or unscrewing using an angular position controller. The jet was driven by a geared down d.c. motor and the angular position measured by a ten turn potentiometer. The controller used reed relays to provide a forward, off, or reverse signal to the motor. This servo provided some control of the fuel air ratio.

4.1.6 Torque Transducer

The existing transducer was a spring balance opposing the rotary motion of the dynanometer and a potentiometer connected to the needle of the balance dial gave an electrical measurement of torque. The dynamic performance of this arrangement was inadequate, attenuating frequencies of the order of 0·03 Hz and above, mainly due to the large inertia of the dynanometer stator. A force balance replacing the spring balance was considered inadequate as excessive forcing from an actuator would have been necessary to give a satisfactory dynamic performance.

An alternative approach used strain gauges and slip rings attached to the shaft coupling the dynanometer to the engine. To avoid overstressing by the large transient torques which occurred when the engine stopped or started, the strain gauges operated at a low-working strain. The strain gauges were connected in a Wheatstone Bridge configuration and, initially, the out of balance voltage from the bridge was amplified with a d.c. amplifier while slip rings transmitted the power supplies and the signal. Slip ring noise and large drifts made this configuration unuseable. The drift was eliminated by using an a.c. bridge supply and a.c. amplification but slip ring noise limited the application of this arrangement.

The effect of slip ring noise was reduced by using a frequency modulated system (A2.2) with its superior noise rejection properties and by regulating and smoothing the power supplies on the shaft using Zener diodes and capacitors. Very careful design of the frequency modulated system was essential to obtain temperature stability as the temperature range encountered was 15°C to 85°C and the operating strain was extremely low. A monostable (A2.3) followed

by an analogue low pass filter demodulated the transducer output to make the device compatible with the computer interface. The final transducer had a flat response up to 15 Hz, 1% linearity and insignificant drift.

4.1.7 Safety and Switching Circuits

Some controls had to be operated by the computer or manually from the control cell or the engine test cell (fig. 4.1.1). The choke was fitted with a solenoid so that it could be operated remotely and the engine dynamo had remote switching in the control cell. The control of the ignition and starter could be switched from the control cell to the computer or the engine test cell. Ignition and starter connections were made with relays providing protection against mains or power supply failure and the absence of cooling water or compressed air and a latching emergency stop control which could only be reset by turning off the main power supply.

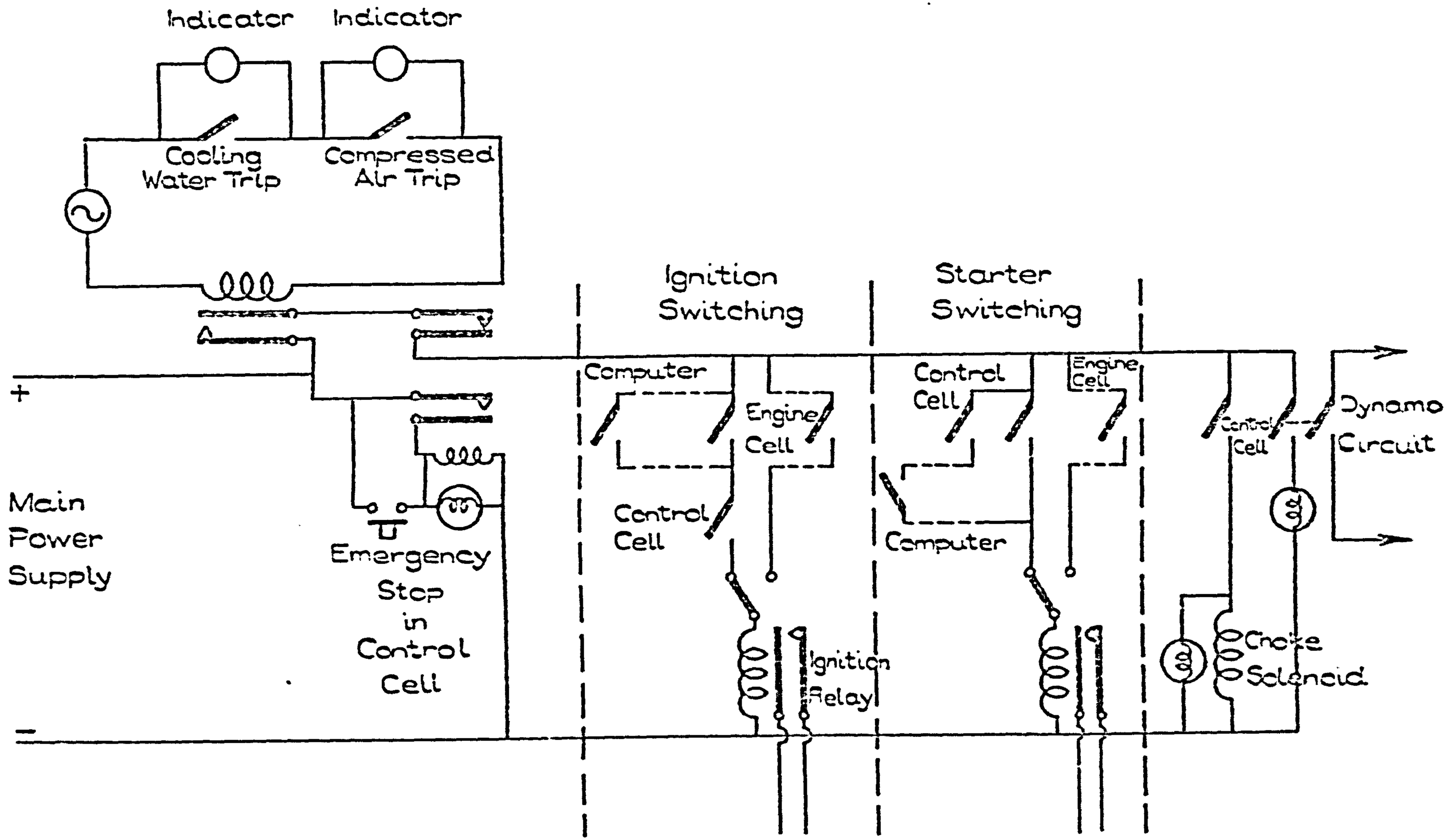


Fig. 4.11. SWITCHING CIRCUITS FOR ENGINE OPERATION.

4.2 Modelling and Control of the Test Rig

4.2.1

A model of the engine rig was evolved to clarify and suggest solutions to the control problems. Initially, a model was developed to represent the system under steady state conditions about a chosen operating point in the middle of the torque and speed ranges. Energy storage elements and time delays were then added to extend the model to represent the system under dynamic conditions and the magnitudes and interactions of these components were evaluated from structural considerations, direct measurement and dynamic tests. The model was developed in terms of a lumped electrical analogy, an established and convenient form for representing the interconnections between the various physical components of a system. Simple controllers were then evaluated using a linearised version of the model programmed for simulation on an analogue computer. More detailed controllers are now being developed by other workers in this field.

4.2.2 Definition of the electro-mechanical analogy

The basis of the analogy is the formal similarity of the mathematical equations describing the properties of the two physically diverse systems. The analogy is defined in table 4.2.1. The choice of current to represent torque and voltage to represent angular velocity was arbitrary and could have been reversed, in which case an inductance represents a moment of inertia and a capacitance represents flexibility.

Table 4.2.1 - Definition of the analogy

Mechanical Quantity	Electrical Quantity	Mechanical Units	Electrical Units
Torque	Current	Poundal Feet	Amps
Angular Velocity	Voltage	Radians/Second	Volts
Inertia	Capacitance	$lbm.ft^2$	Farads
Viscous Damping	Resistance	Rads/sec.ft.pdl.	Ohms
Compliance	Inductance	Rads/pdl.ft.	Henrys

4.2.3 Steady-state model of the test rig

Difficulty in determining experimental points for the static characteristics of the rig, in the low speed high torque region, due to the unstable operation was alleviated by using proportional speed control acting on the dynamometer field current. The engine torque/speed curves obtained for the full range of throttle servo inputs are given in fig. 4.2.1 and the dynamometer torque/speed curves for a range of excitation current are given in fig. 4.2.4. These show that a small signal linear model is applicable over a large range of operation providing that the low speed high torque region is excluded.

The engine was represented by a voltage source V_E with internal resistance R_E shown in fig. 4.2.2a, R_E being given by the tangent to the torque speed curve at the chosen operating point and V_E by the intercept of this tangent with the torque axis. V_E was a function of throttle opening and the relationship between V_E and the input voltage to the throttle servo V_T is shown in fig. 4.2.3. The small signal model for the dynamometer, fig. 4.2.2b, was given by a current generator I_D in parallel with a resistance R_D, the component

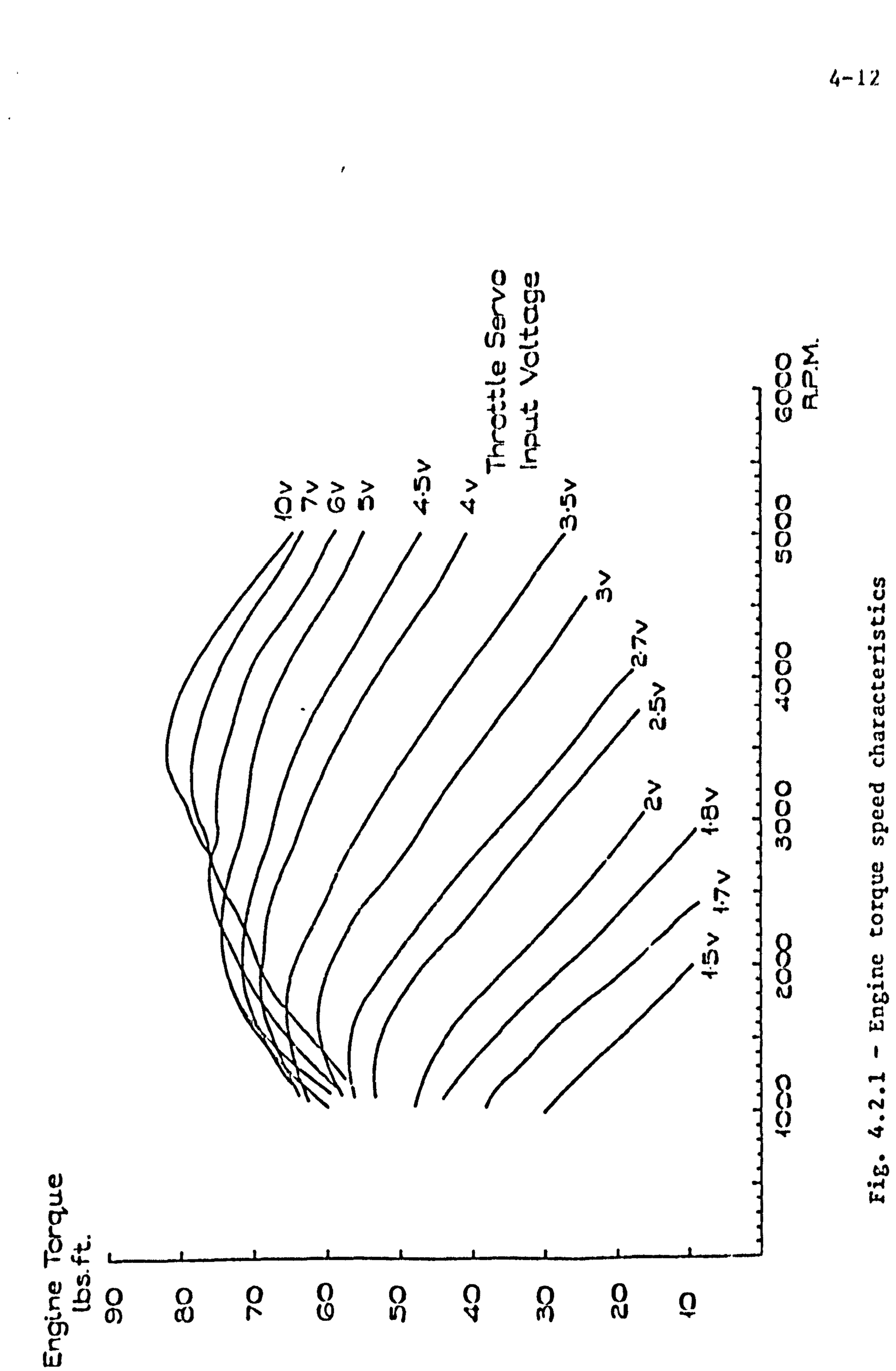


Fig. 4.2.1 - Engine torque speed characteristics

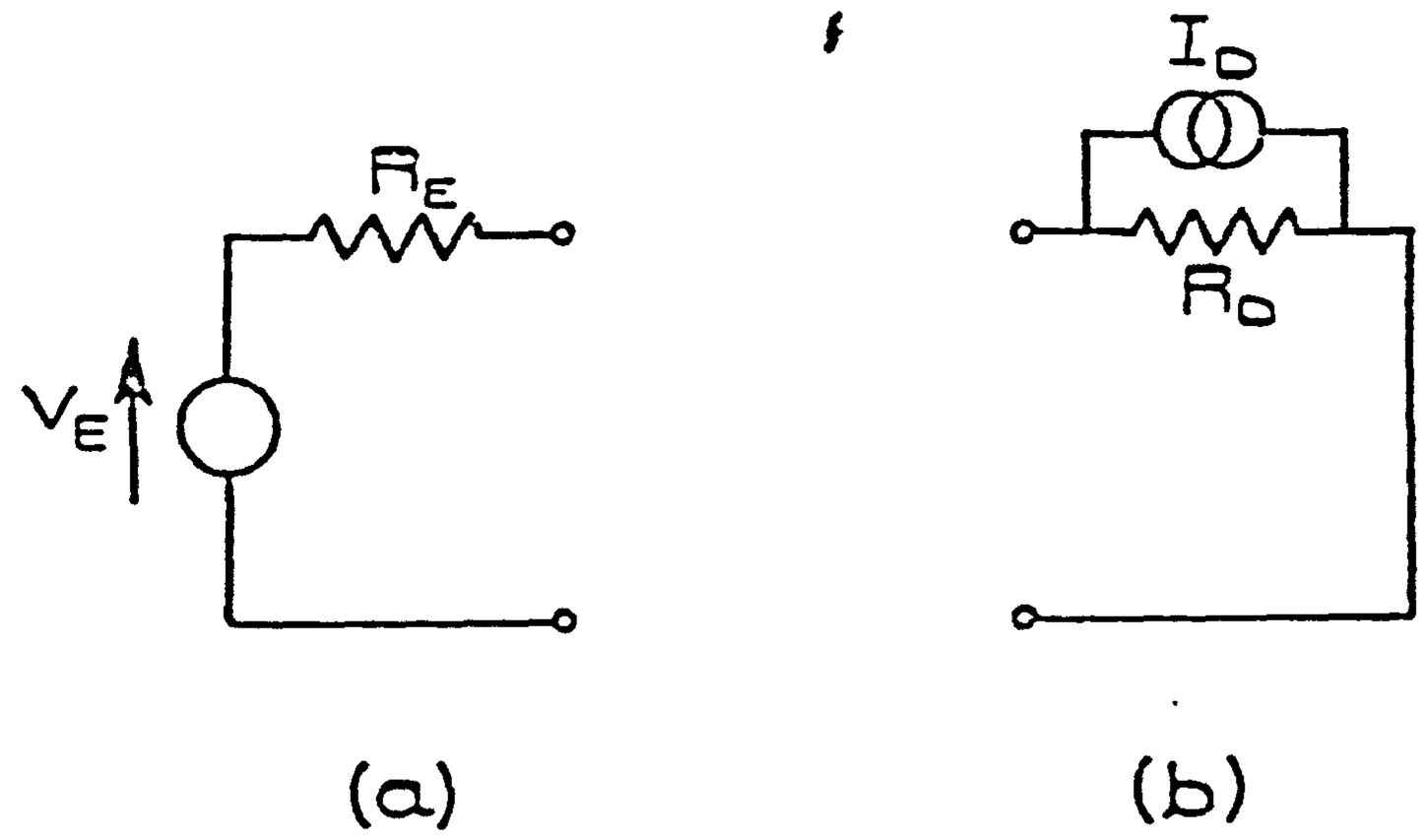


Fig. 4.2.2 - Steady state models of (a) engine (b) dynamometer

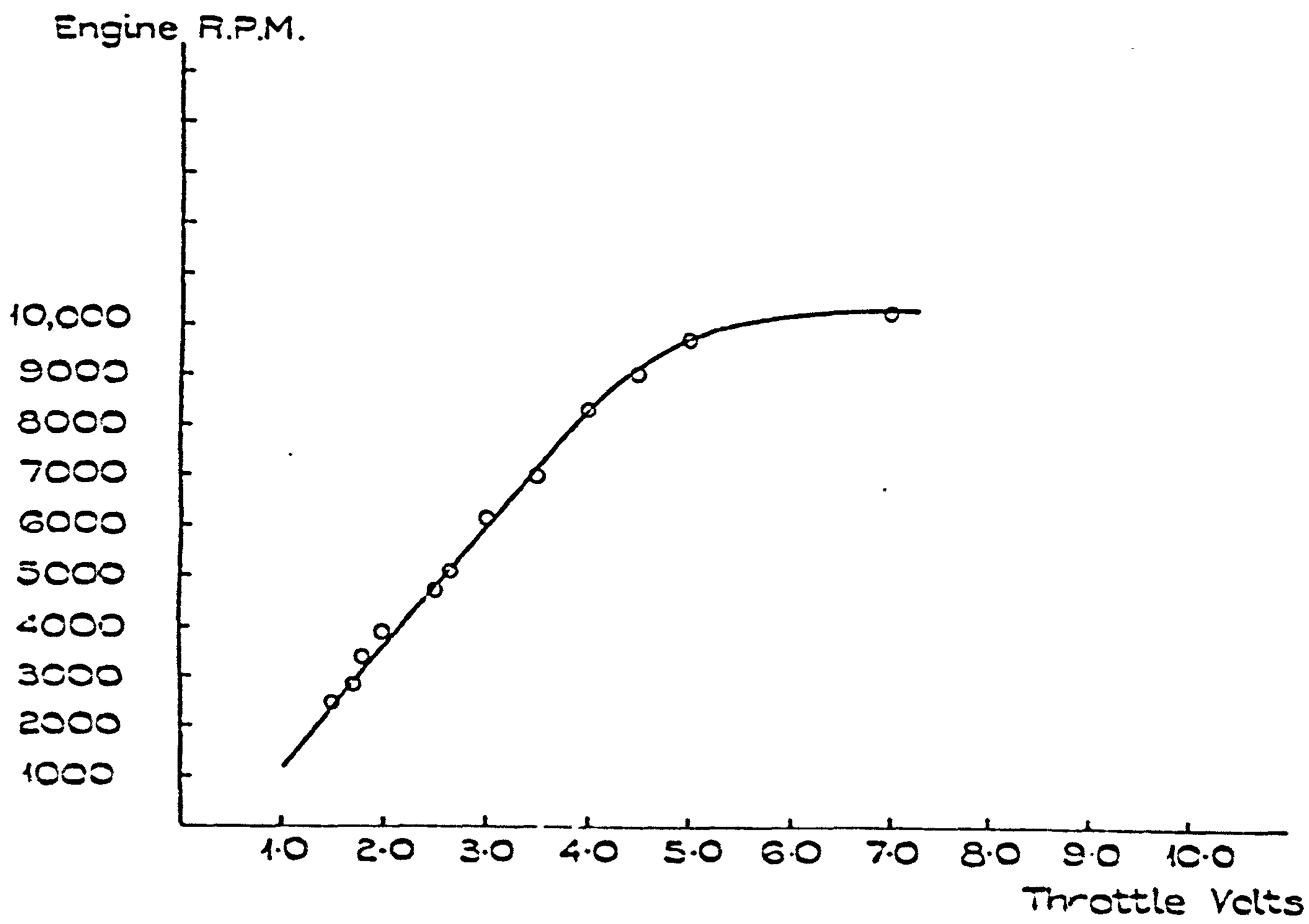


Fig. 4.2.3 - Throttle steady state transfer characteristic

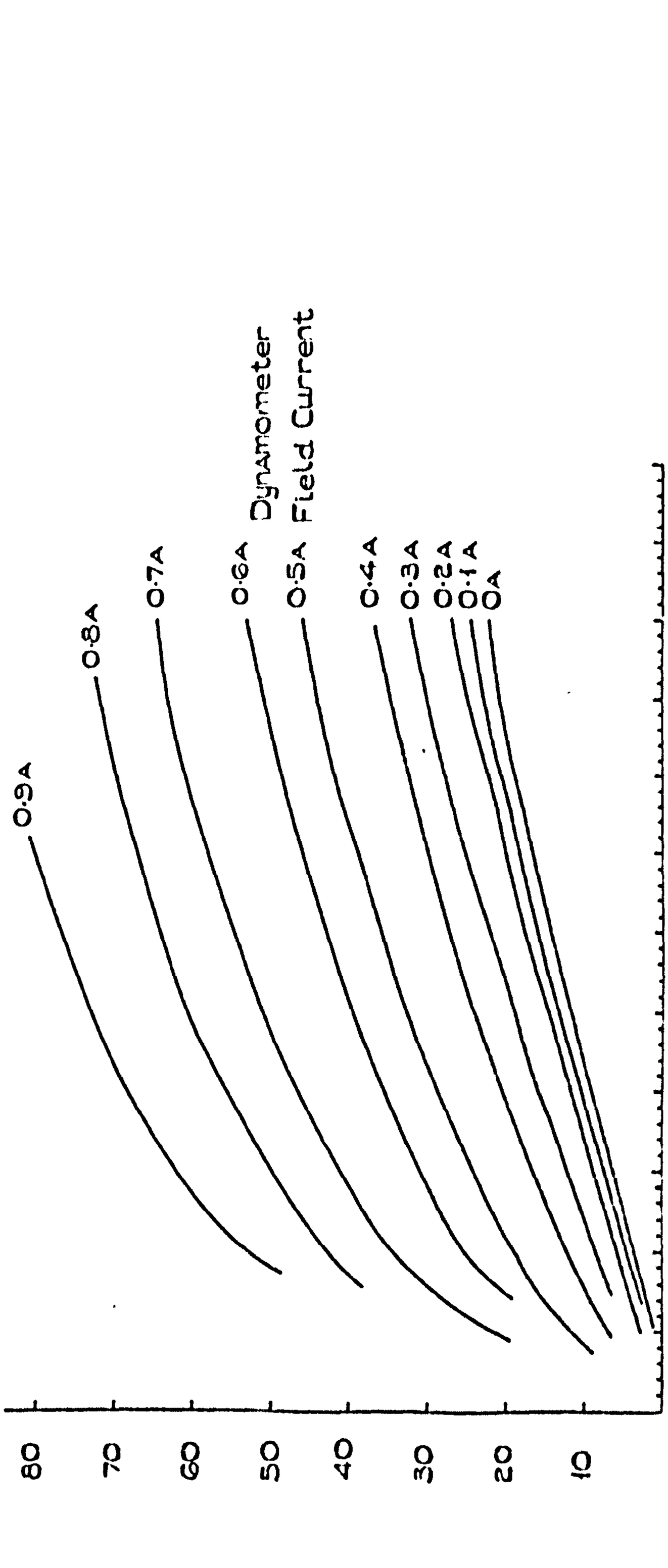


Fig. 4.2.4 - Dynamometer torque-speed characteristic

values being obtained in a similar manner to those of the engine. The relationship between I_D and the dynamometer field current I_F is shown in fig. 4.2.5. In the steady state model, the coupling shaft was represented by a short circuit as no velocity difference can occur across it when the transients have died away.

4.2.4 Extension of the model to include dynamics

Consideration of the detailed structure of the engine would have led to a complex model which included energy dissipation elements to represent the service, pumping, thermodynamic and mechanical losses and energy storage elements to represent the mass, inertia and compliance of the various reciprocating and rotating parts. Instead a simpler model fig. 4.2.7a was developed by obtaining the most significant dynamics of the engine experimentally. The engine flywheel was decoupled from the shaft connecting it to the dynamometer and a series of transfer functions between the input voltage to the throttle servo and the engine speed obtained over a range of mean speeds using a digital transfer function analyser (T.F.A.)[20]. A typical result is shown in fig. 4.2.6. Examination of the amplitude plots of the transfer function, suggested a first order model and the phase plots corresponded to a first order system with a time delay. This time delay was found to be inversely proportional to the engine speed and was accounted for by the average time taken to inhale and ignite a charge. Finally the capacitor C_E was introduced into the model to represent the lumped inertia of the engine, its value being derived from the engine source resistance R_E and the time constant of the first order system.

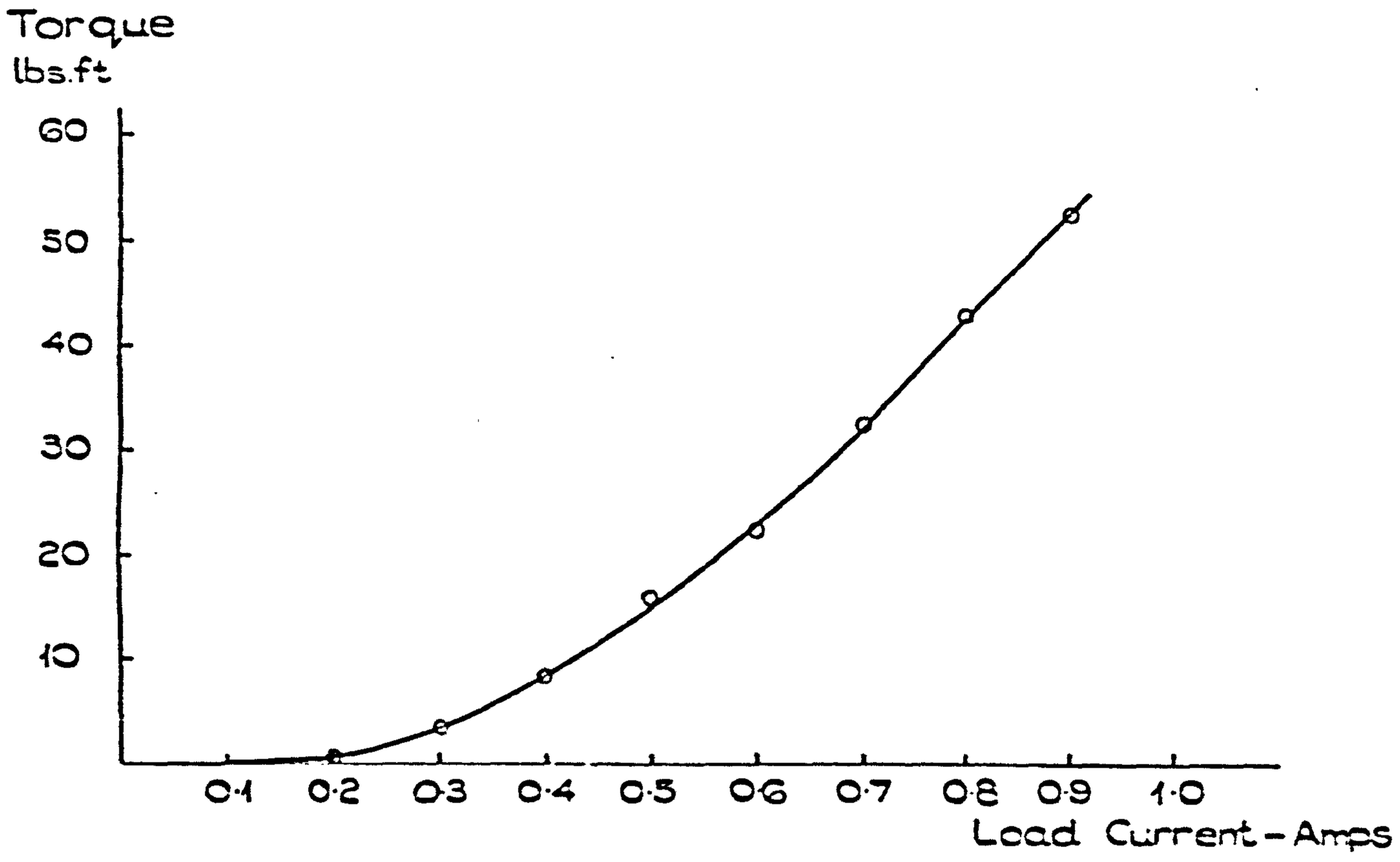
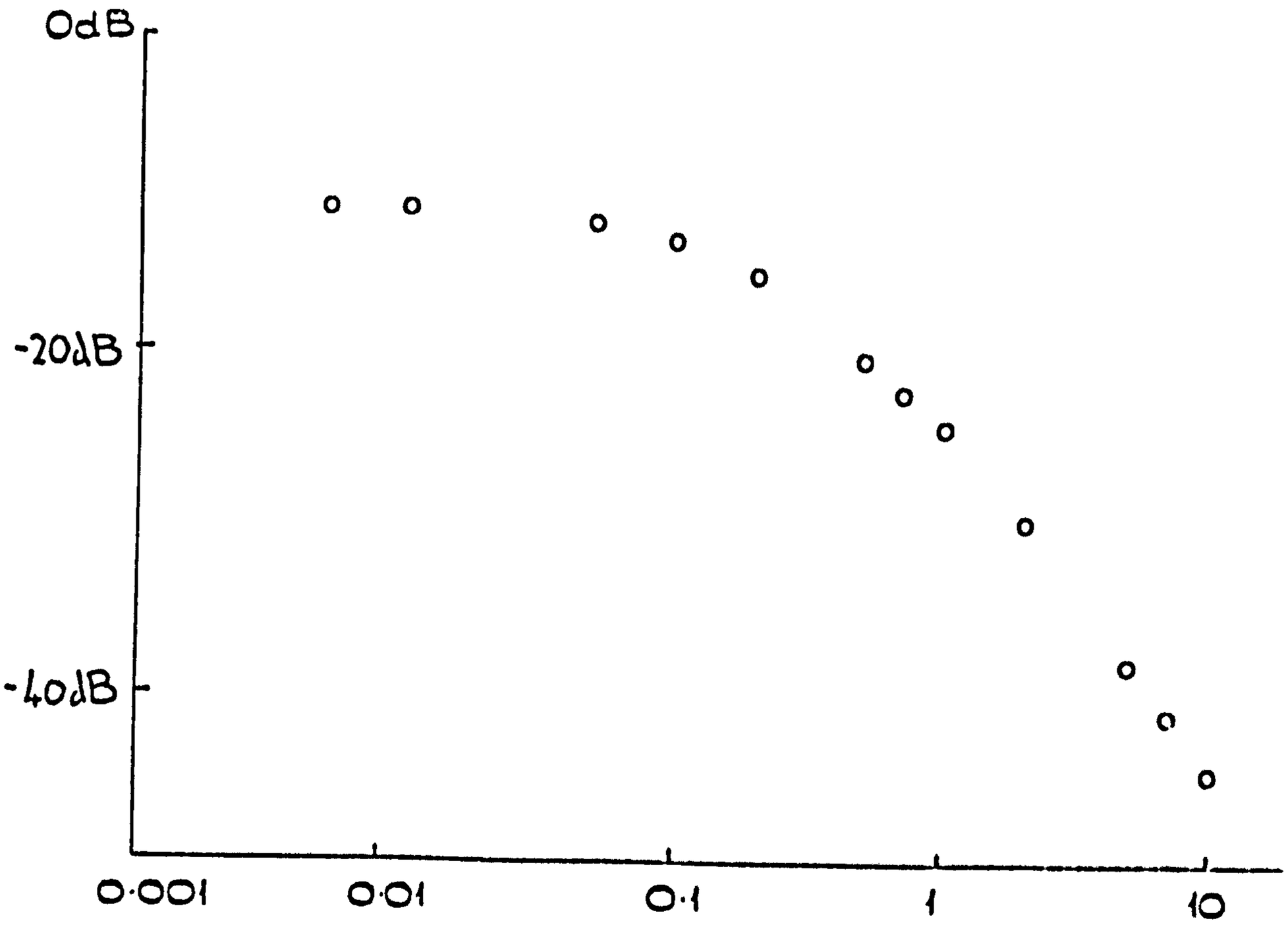


Fig. 4.2.5 - Dynamometer steady state transfer characteristic



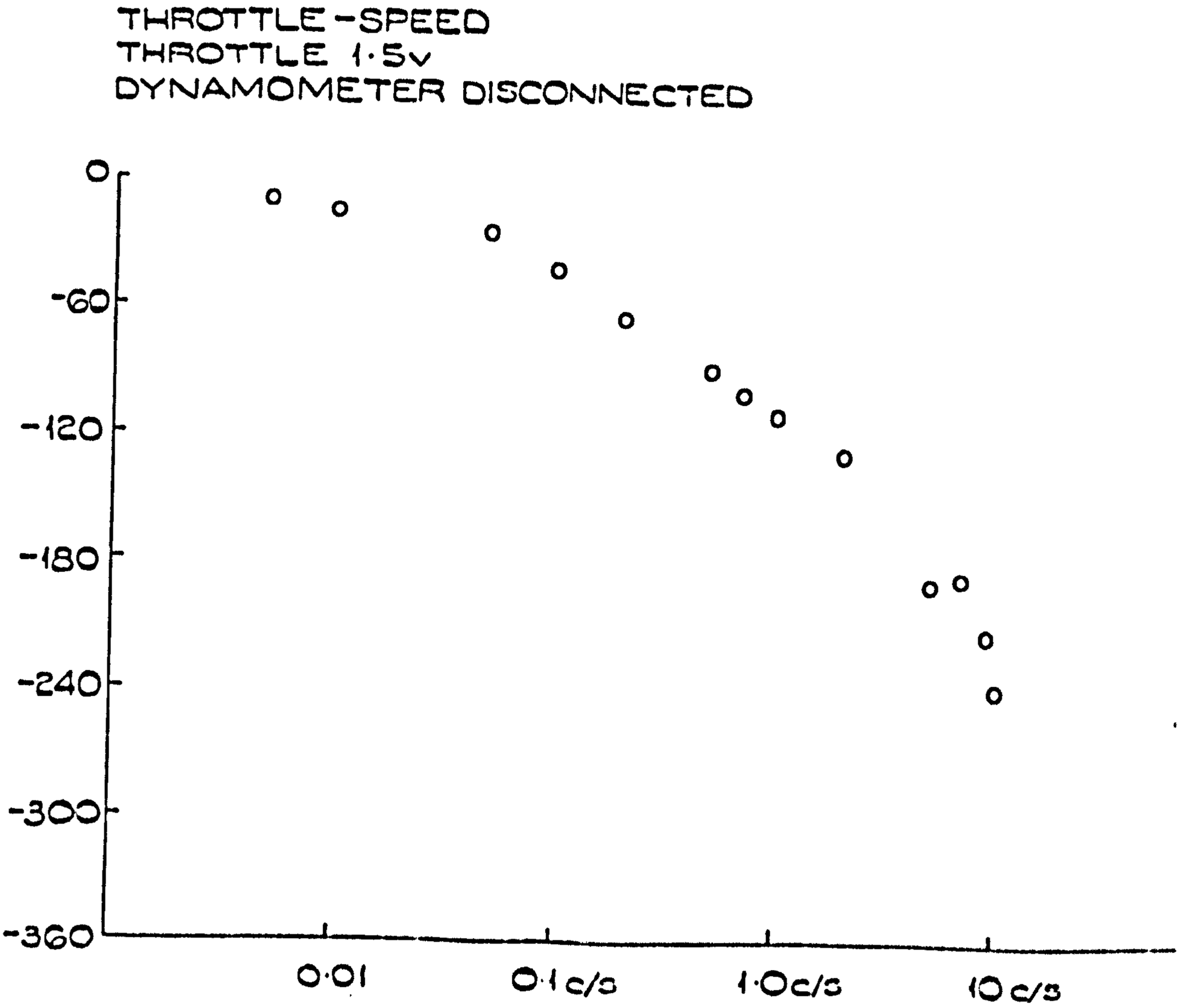


Fig. 4.2.6 - Throttle-speed frequency response of the engine

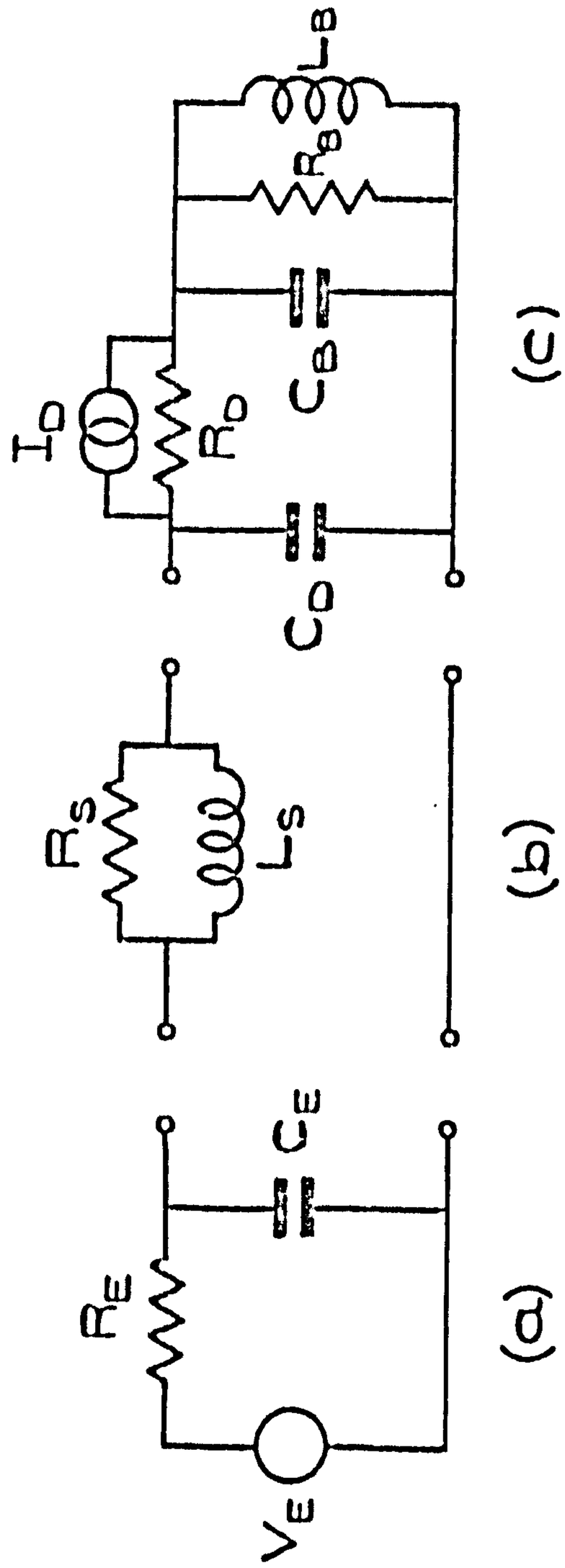


Fig. 4.2.7 - Basic dynamic model at (a) engine (b) shaft (c) dynamometer

The dynamic model of the dynamometer is given in fig. 4.2.7c. The dynamometer possessed a large rotor inertia evaluated by a simple mechanical test and giving the value of the capacitance C_D. The resistance R_D was introduced to account for the power dissipation which was necessarily taking place because a speed difference existed between the dynamometer rotor and its casing while torque was transmitted between them. The outer casing of the dynamometer was connected through a lever arm to weighing gear, consisting of a spring and dashpot system anchored to the inertial frame of reference. The component values associated with the dynamometer outer casing and weighing gear, the capacitance C_B, the resistance R_B and the inductance L_B, were evaluated from direct measurement of the spring and lever arm, and a step response of the balance system.

Finally, the connecting shaft was represented by an inductor in parallel with a resistor, as it behaved as a torsional spring with internal damping. The shaft stiffness was measured directly, and the internal damping derived by inspecting the transient response of the anchored shaft attached to the dynamometer rotor. The model of the shaft is shown in fig. 4.2.7b, and table 4.2.2 assigns the equivalent electrical values to the complete model.

Table 4.2.2 - Numerical Values for the electrical analogy

R_E	1·19Ω	
C_E	0·447F	
L_S	5·32 × $10^{-4}H$	1 volt ≡ 52·4 rads/sec
R_S	0·455Ω	(10v ≡ 500 r.p.m.)
C_D	1·38F	
R_D	3·00Ω	1 amp ≡ 322 pdl.ft
C_B	1·57 × $10^{3}F$	(10v ≡ 100 lbf.ft)
R_B	0·154 × $10^{-3}\Omega$	
L_B	0·112 × $10^{-4}H$	

4.2.5 Experimental verification of the model

The dynamometer casing was originally clamped to the inertial frame of reference to reduce the order of the model and simplify the problem of comparing the model with the plant. Subsequently, this operation also permitted the design of closed loop torque and speed control systems with a marginally greater bandwidth than would otherwise have been possible. This simplified form of the model is shown in fig. 4.2.8.

The initial verification of the model was carried out using sine-wave testing techniques. Subsequently on-line techniques using pseudo-random binary sequences (p.r.b.s.) were employed and finally step responses for the plant were compared with those obtained from the analogue computer model.

The measured variables available were the dynamometer speed, V_D, and the shaft torque, I_S, and the input variables were the voltage applied to the throttle servo, V_T, and the dynamometer field current controller, V_L. The small signal dynamic performance could therefore be defined by four transfer functions. An Algol 60 programme was written to calculate the theoretical frequency responses for the model, and measurements over the range 0·001 to 15 Hz taken with the T.F.A. on the rig. The experimental and theoretical transfer functions between the throttle servo input and dynamometer speed with shaft torque (figs. 4.2.11 to 4.2.12) showed satisfactory agreement, but the responses of the shaft torque and dynamometer speed to dynamometer field control perturbations (figs. 4.2.9 to 4.2.10) were in poor agreement. It was suggested that the extra attenuation at

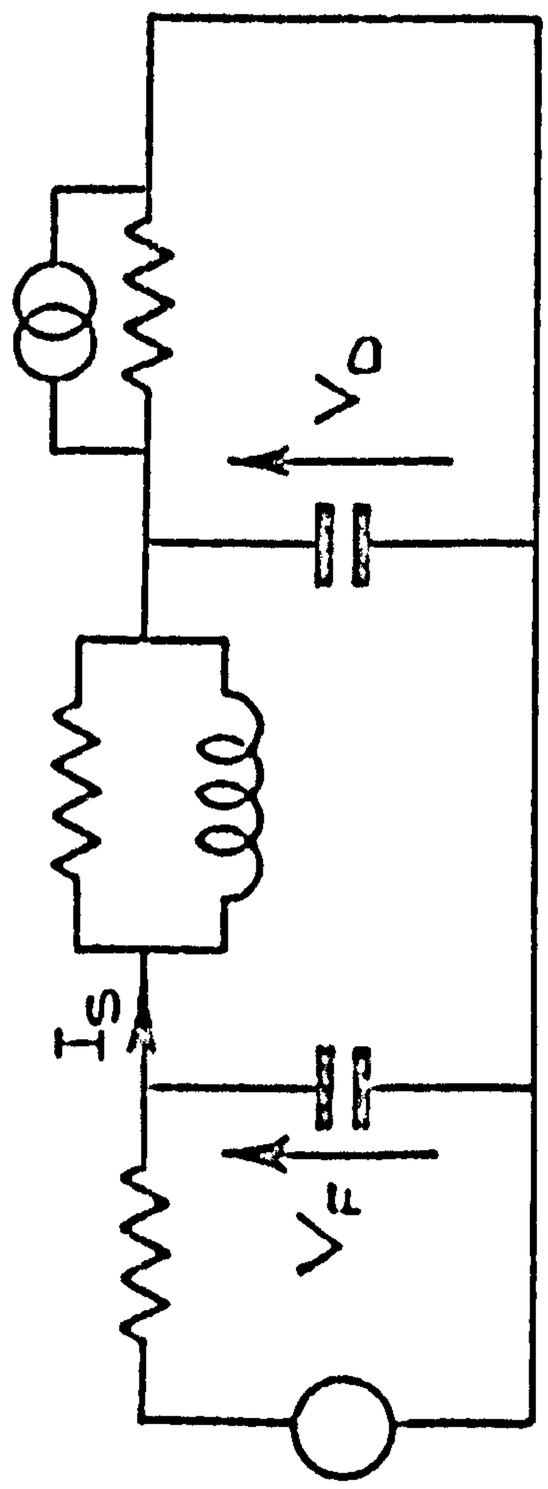


Fig. 4.2.8 - Simplified dynamic model of the rig

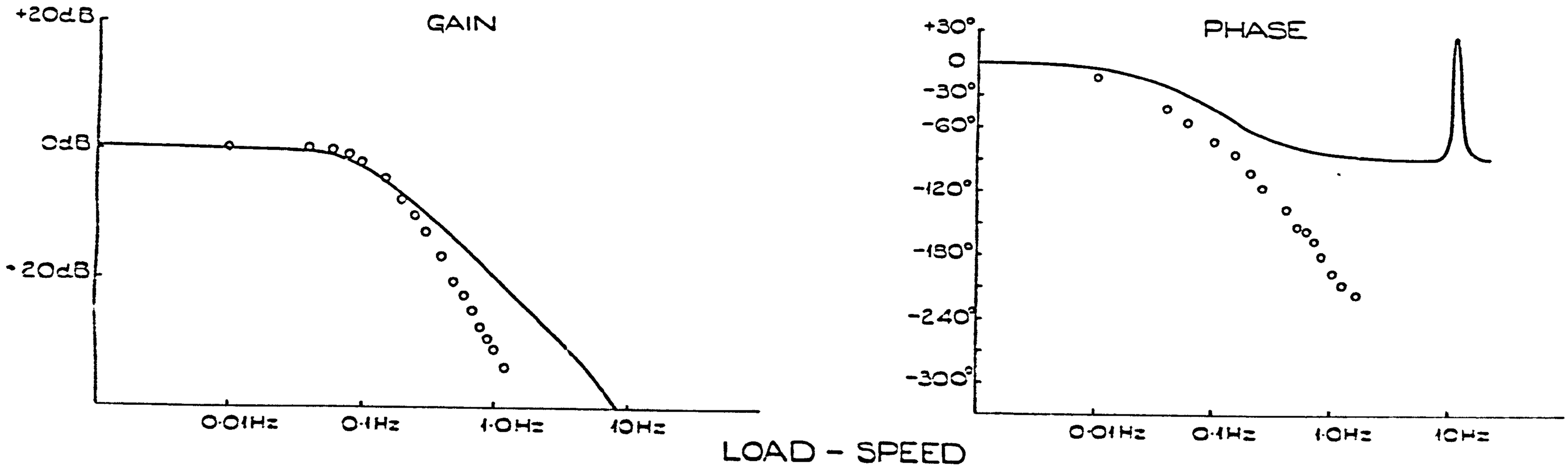


Fig. 4.2.9 - Frequency response of speed to load perturbations
The solid line shows computed results

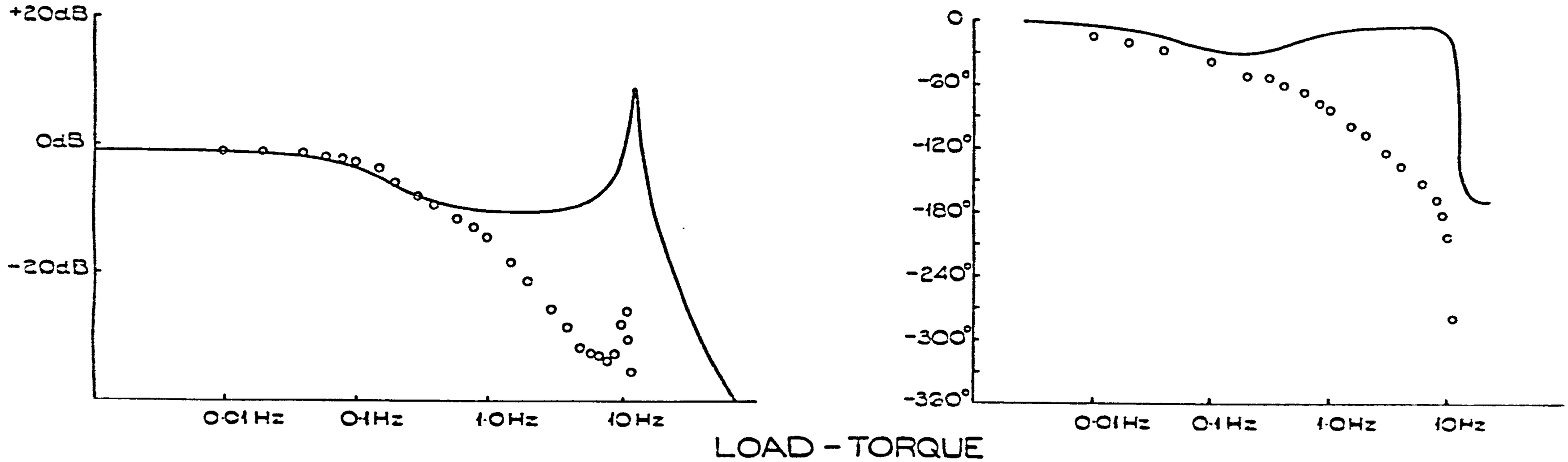


Fig. 4.2.10 - Frequency response of torque to load perturbations

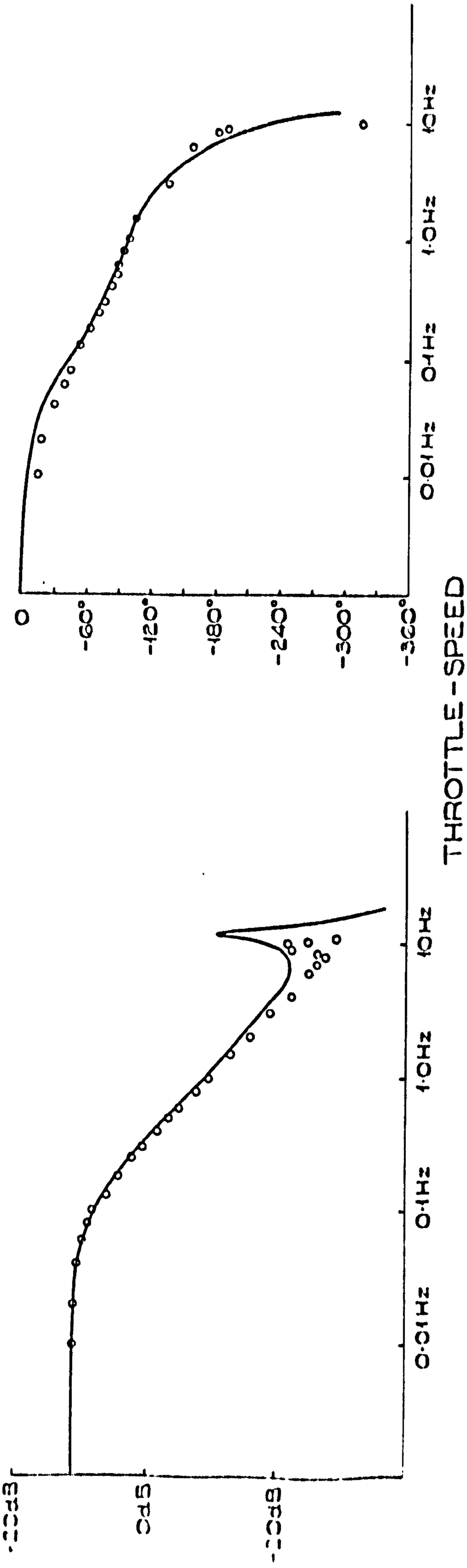


Fig. 4.2.11 - Frequency response of speed to throttle perturbations

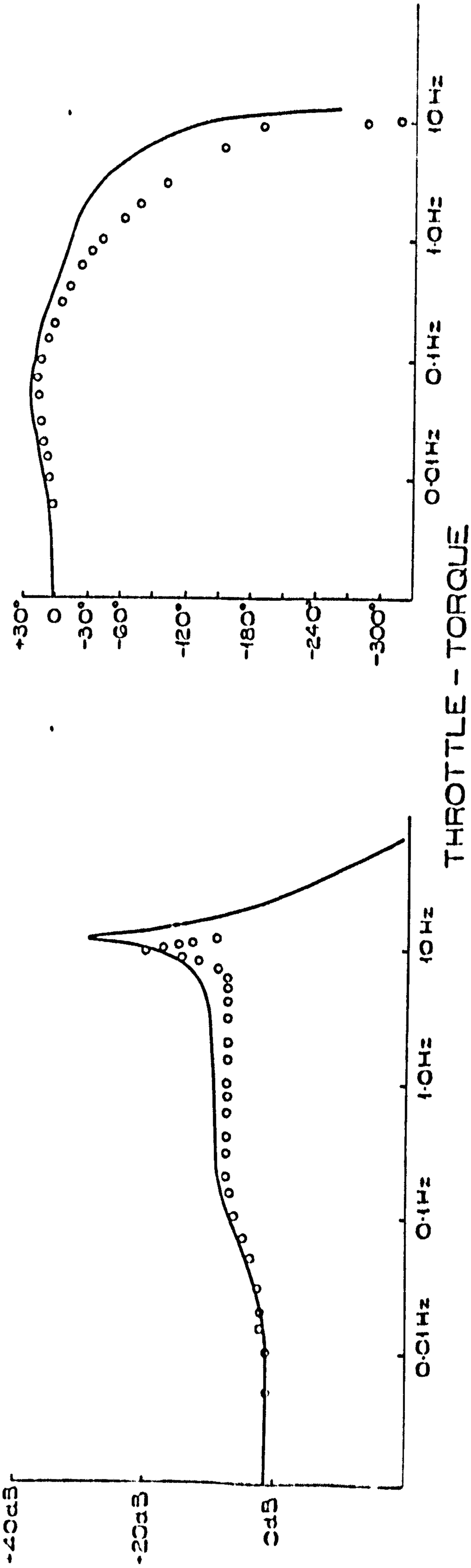


Fig. 4.2.12 - Frequency response of torque to throttle perturbations

high frequencies of the plant was due to the flux build-up associated with the dynamometer casing and rotor, and the frequency response between the excitation current and flux was therefore examined using a search coil. Examination of this response (fig. 4.2.13) suggested that the flux build-up contributed two cascaded first order lags and the revised model including these lags is shown in fig. 4.2.14. The revised frequency responses of the shaft torque and dynamometer speed to the input to the dynamometer field current control (figs. 4.2.15 and 4.2.16) showed considerably better agreement.

Identification using p.r.b.s. was then examined. To ensure a flat power spectrum over the frequency range of interest, the p.r.b.s. bit rate should be made twice the highest frequency concerned within the identification and the period of the sequence should be defined by the slowest time constant. A bit frequency of 25 Hz and a period of 150 sec. corresponded to the range of engine time constants and the nearest available sequence has 4095 bits with a period of 150 secs. and harmonic spacing every 0·0066 Hz. Although this would have given excellent frequency resolution, it required excessive computer time and storage for the computation of all the shifts required to describe the system and, as an alternative, three experiments were carried out using a 31 bit sequence with bit rates of 10 sec., 1 sec. and $\frac{1}{25}$ sec. respectively. The different fundamental frequencies imply different harmonic spacings, the fastest sequence providing a frequency resolution of ·806 Hz, which was marginally satisfactory for identifying the system resonance which has a bandwidth of the same order.

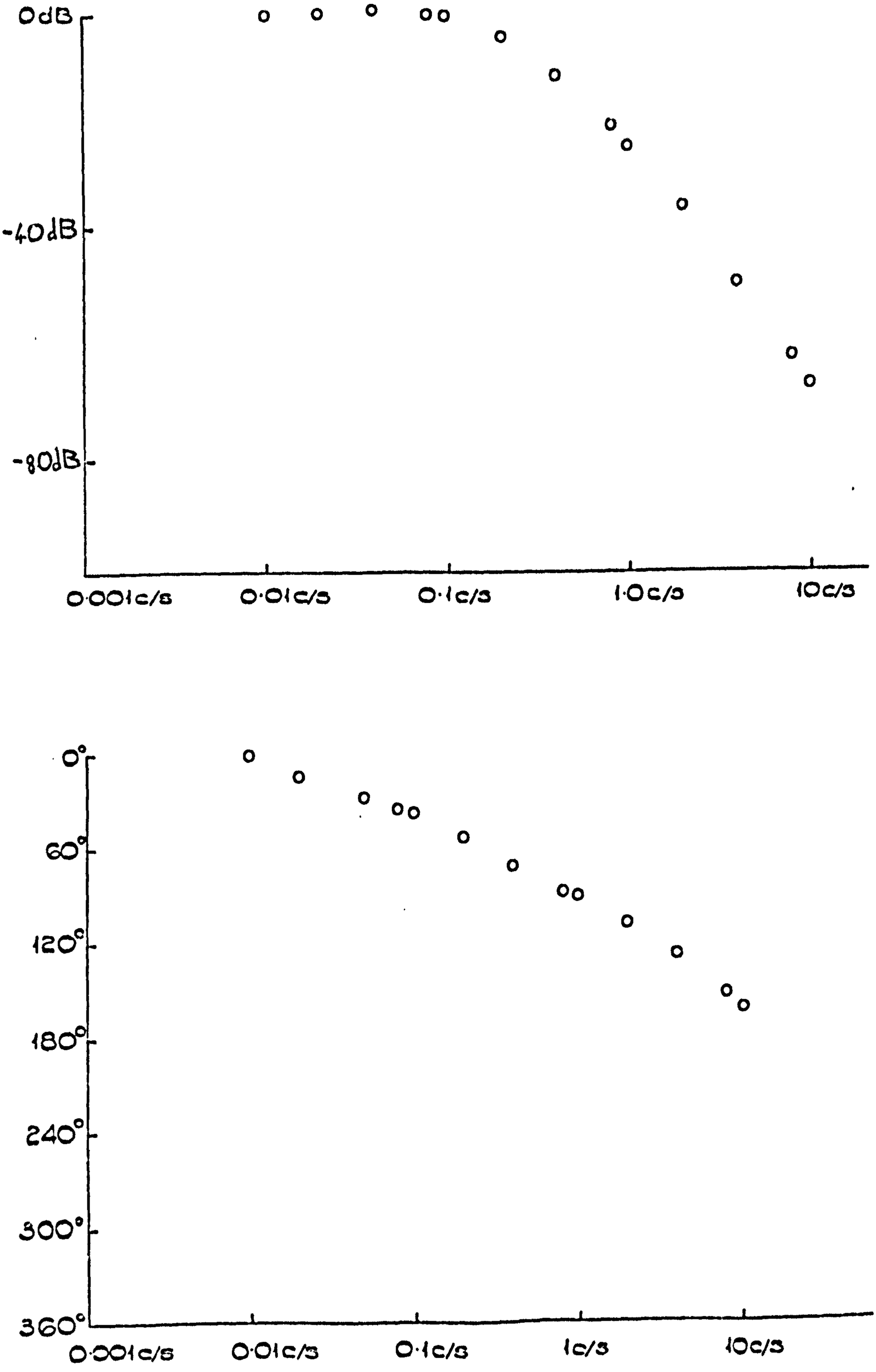


Fig. 4.2.13 - Frequency response of dynamometer flux to field current perturbations

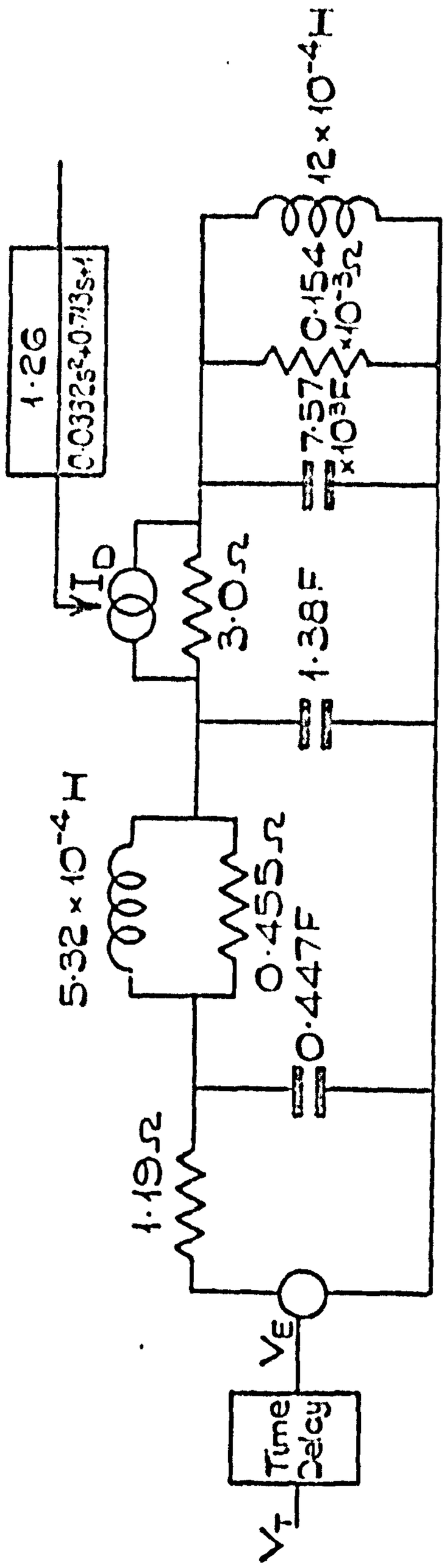


Fig. 4.2.14 - Revised dynamic model

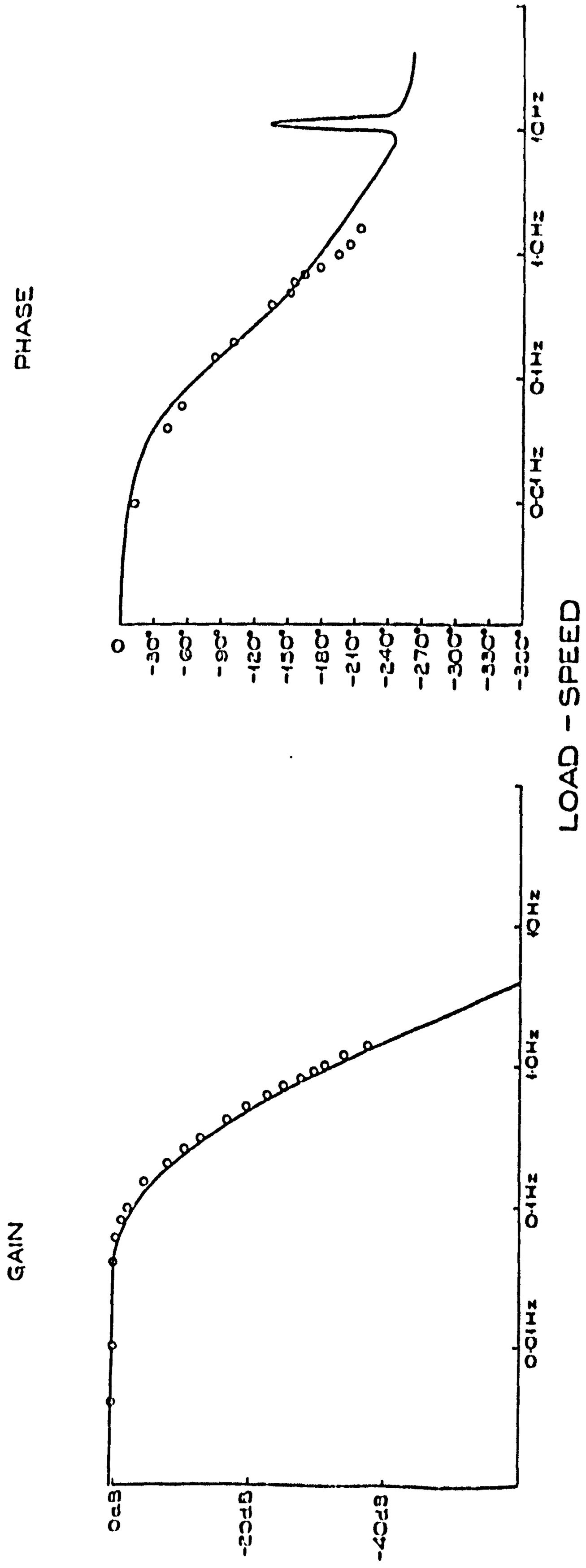


Fig. 4.2.15 - Modified load-speed frequency response

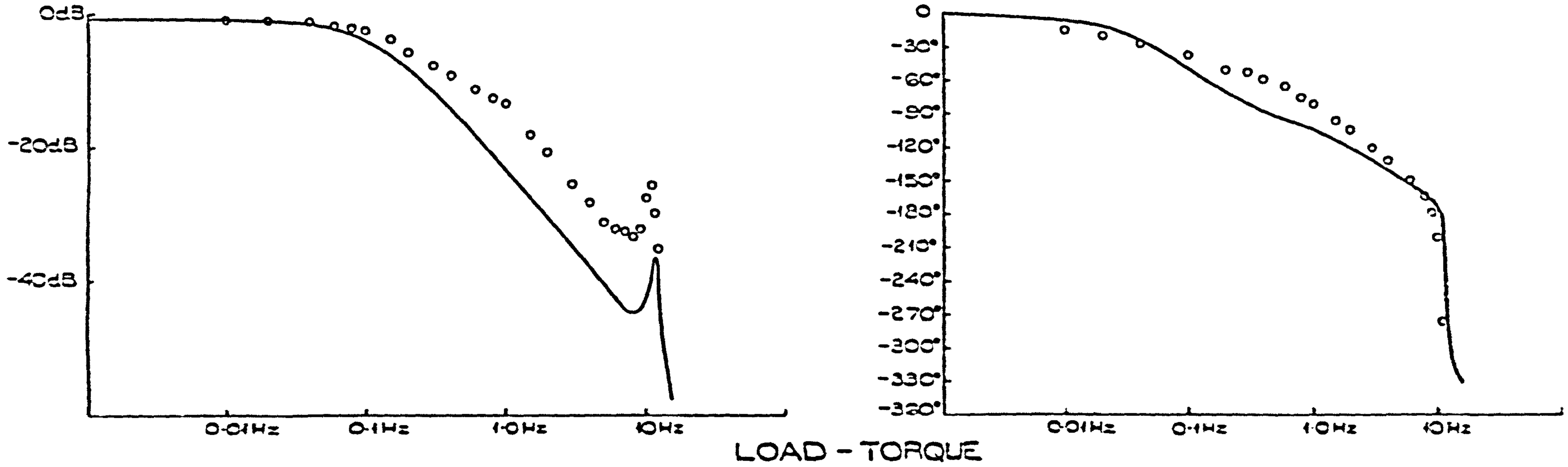


Fig. 4.2.16 - Modified Load-torque frequency response

The impulse responses obtained from the p.r.b.s. perturbations are shown in figs. 4.2.17 and 4.2.18. The effect of the digital filter used to attenuate the high frequencies can be taken into account by assuming a modified auto-correlation function of the p.r.b.s. (fig. 4.2.19). The throttle servo suffered from velocity limiting when following a two-level sequence, with the amplitudes used, significant attenuation occurred at 8 Hz and above. For all other responses the rig heavily attenuated the high frequencies and the swamping of the resulting signals by noise made identification at these frequencies difficult. A more satisfactory response for engine perturbation was obtained by perturbing the ignition setting in place of the throttle and the result is shown in fig. 4.2.20.

The step responses for the plant and analogue computer model are shown in figs. 4.2.21 and 4.2.22. It was difficult to identify the high frequency performance of the rig from these response because the throttle velocity limit significantly affected the high frequency components of a large step input.

4.2.6 Dynamic equations and analogue computer simulation

The three state variables chosen for the electromechanical model were the engine speed, V_E, the dynamometer speed, V_D, and the shaft torque, I_S. The state space matrix equations for the system, assuming linear operation, are:

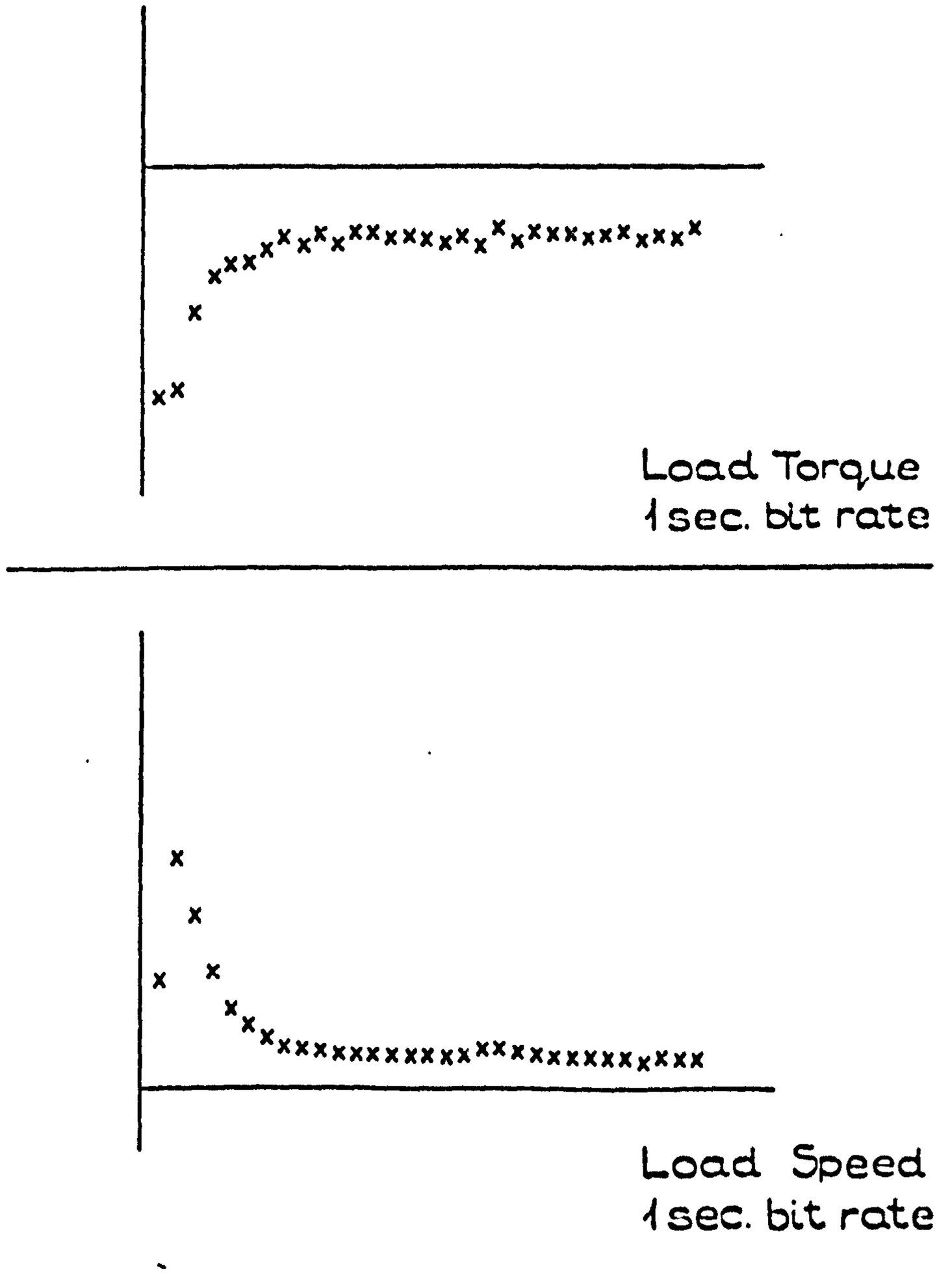


Fig. 4.2.17 - Estimates of impulse responses using p.r.b.s. perturbations on the load

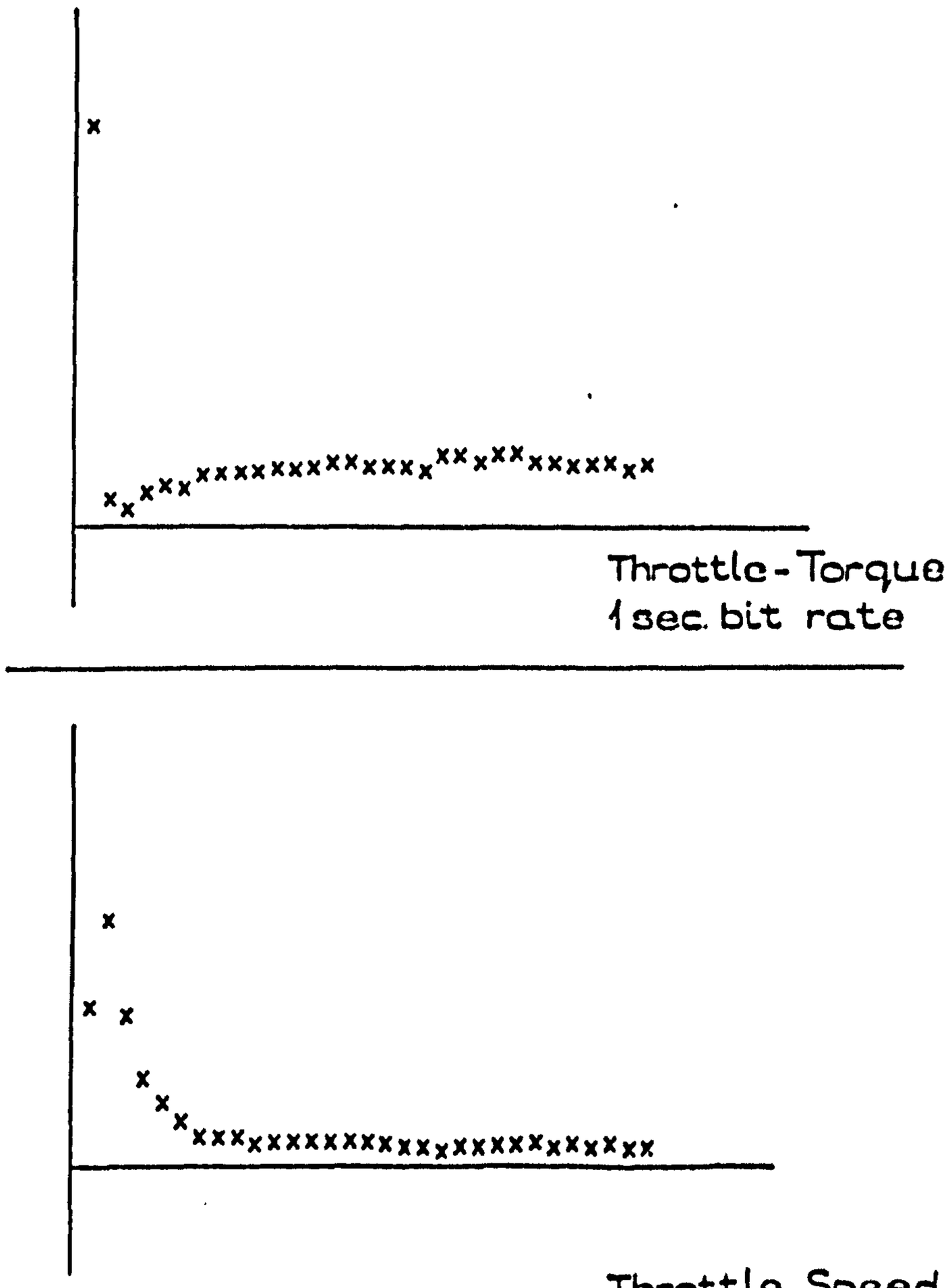


Fig. 4.2.18 - Estimate of the impulse responses using p.r.b.s. perturbations on the throttle

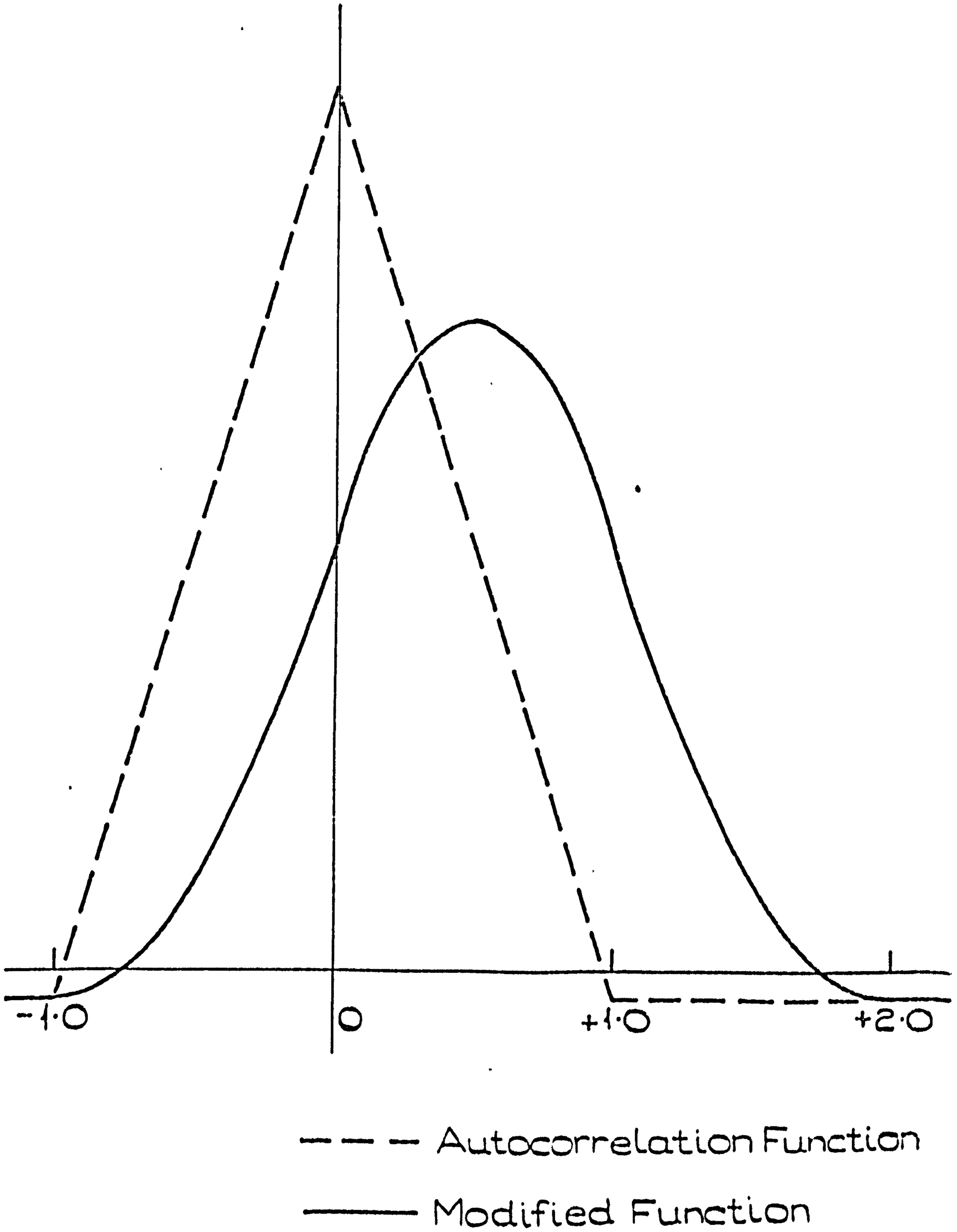


Fig. 4.2.19 - Effect of the running average filter on the autocorrelation function of a p.r.b.s.

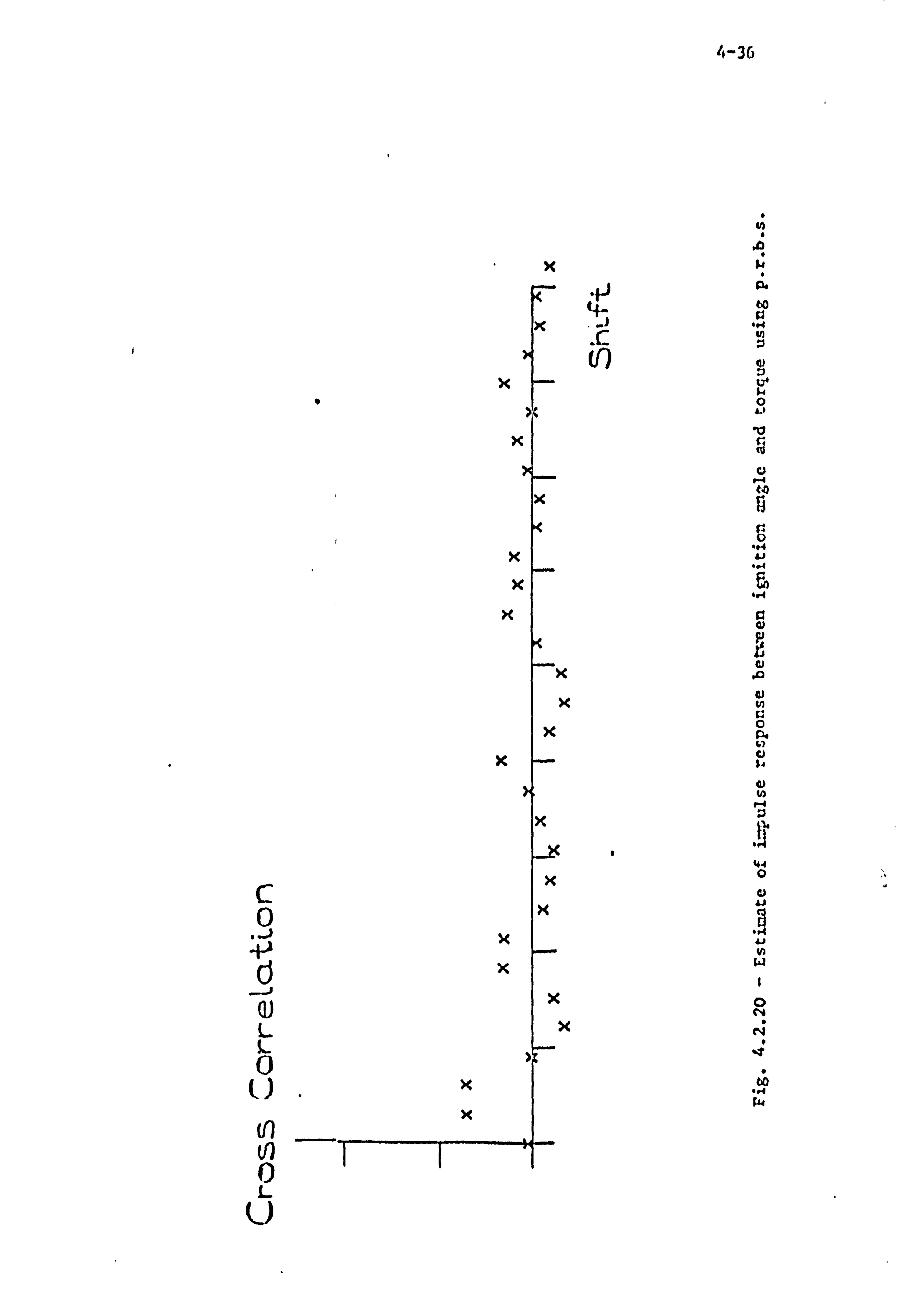


Fig. 4.2.20 - Estimate of impulse response between ignition angle and torque using p.r.b.s.

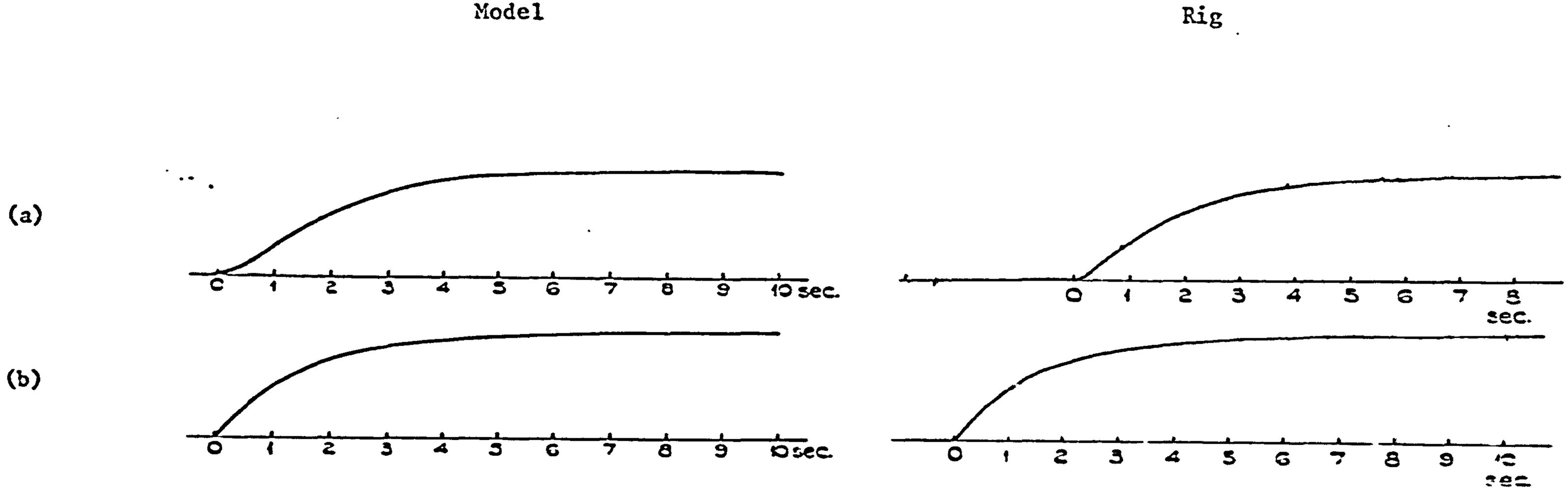


Fig. 4.2.21 - Step responses from the rig and model (a) load-speed (b) throttle-speed

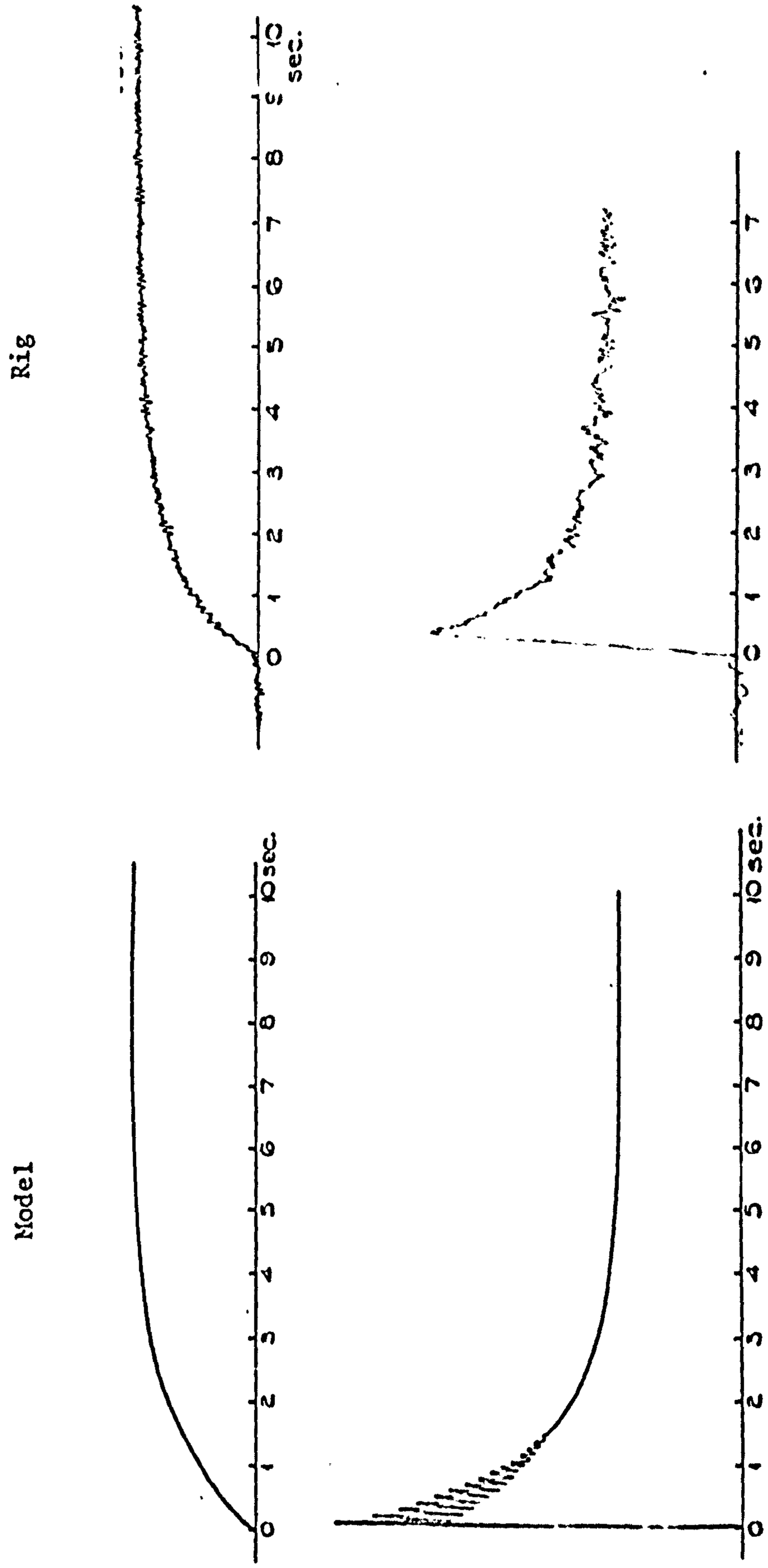


Fig. 4.2.22 - Step Responses from the rig and model (a) load-torque (b) throttle-torque

$$\begin{bmatrix} \frac{dV_F}{dt} \\ \frac{dV_D}{dt} \\ \frac{dI_S}{dt} \end{bmatrix} = \begin{bmatrix} -\frac{1}{C_E R_E} & 0 & -\frac{1}{C_E} \\ 0 & -\frac{1}{C_D R_D} & \frac{1}{C_D} \\ \left(\frac{1}{L_S} - \frac{1}{C_E R_E R_S}\right) & \left(\frac{1}{C_D R_D R_S} - \frac{1}{L_S}\right) & -\left(\frac{1}{C_E R_S} + \frac{1}{C_D R_S}\right) \end{bmatrix} \begin{bmatrix} V_F \\ V_D \\ I_S \end{bmatrix}$$

$$+ \begin{bmatrix} \frac{1}{R_E C_E} & 0 \\ 0 & -\frac{1}{C_D} \\ \frac{1}{C_E R_E R_S} & \frac{1}{C_D R_S} \end{bmatrix} \begin{bmatrix} V_E \\ I_D \end{bmatrix},$$

where $V_E = 4.86 V_T$,

and $0{\cdot}0332 \frac{d^2 I_D}{dt^2} + 0{\cdot}713 \frac{dI_D}{dt} + I_D = 1{\cdot}26 V_L$

The analogue computer diagram for the above equations is given in fig. 4.2.23.

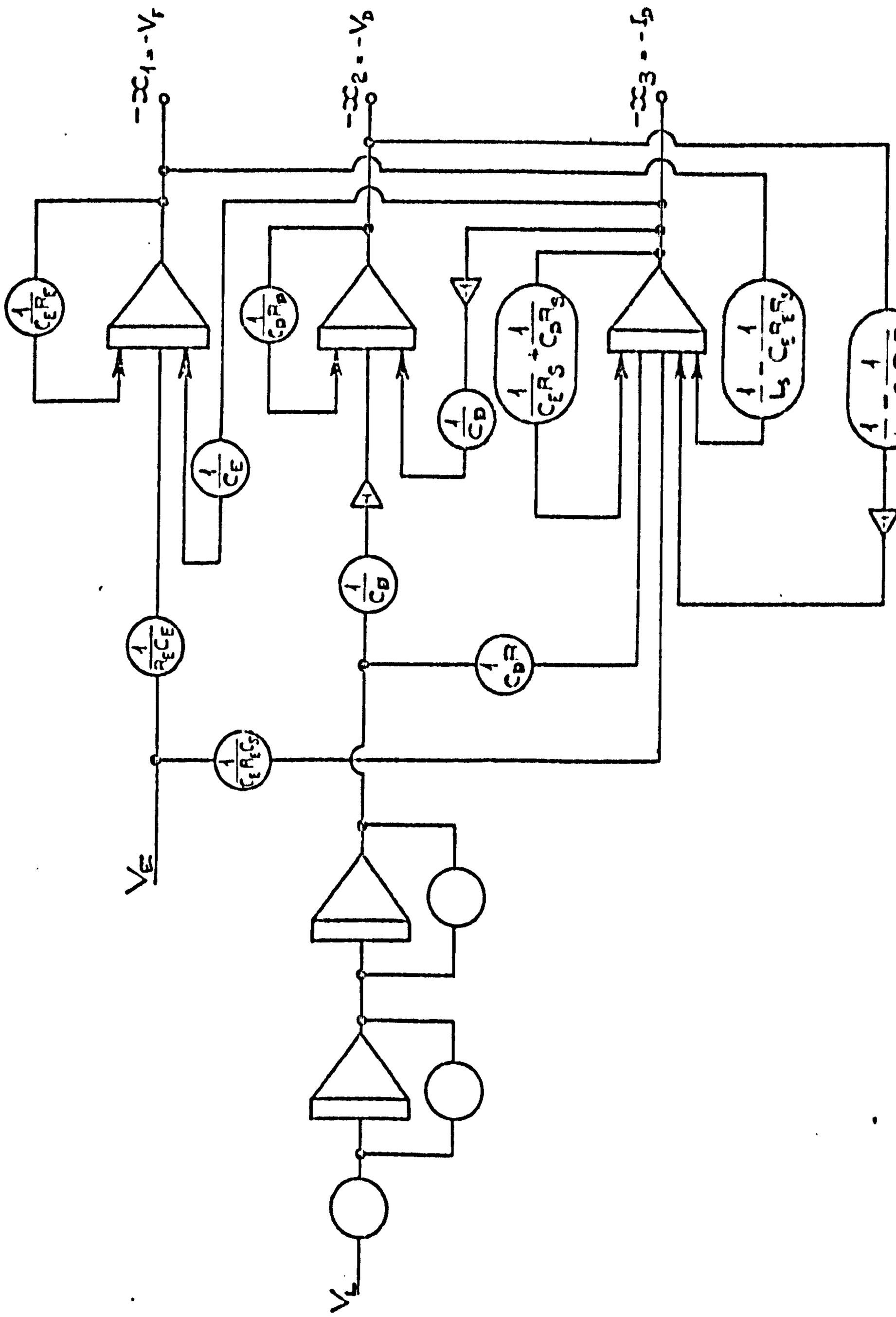


Fig. 4.2.23 - Analogue computer model

4.3 Engine Operating Programme

4.3.1 Programme Specification

This programme has been designed to enable the computer to communicate with its external environment through a variety of channels. The time scale of operation is critical and may be determined by the external environment. Subprogrammes must be organised in such a way that the machine can respond within a suitable response time to the external or timing demands. This has been achieved by working on a priority basis and, where working times were known, adjustments made to the programme to alleviate any congestion. If a computation requires data from different instants for its completion, it is preferable to calculate intermediate results on receipt of information so as to spread computation time demands as evenly as possible.

The basic structure consists of an overall executive, executive directives, engine directives, engine subprogramme, and test tapes. The executive has been written in general terms for use with most small processes and the executive directives will be found useful for most on-line systems. The engine directives are specialised programmes for the particular engine rig used, the engine subprogrammes provide special functions for a particular test, and finally, test tapes give the sequence of operations.

4.3.2 Executive

On-line real-time computer operation of external processes requires a timing system, data input and output of the process variables, data input and access by the operator, servicing and direction of the interrupt system and provision for background computation. In most systems, it will also be necessary to have some form of limit checking and alarm facilities.

The computer has an internal clock, operating at 50 Hz, which generates an interrupt with the second highest priority. For the operation of the test rig, other timings must be generated, down to the one per minute necessary for noting experiment times. These other clocks are used to generate software interrupts after the executive demands at a particular instant have been completed. Alternative 50 Hz clock interrupts are used to generate the 25 Hz software interrupt and the remainder generate five 5 Hz software interrupts sequentially, the fifth being further divided to yield the lower frequencies. (fig. 4.3.1)

The executive processes a group of four analogue inputs. (fig. 4.3.2) The addresses of the four current input channels are obtained from an allocated block of storage. This also contains a flag to indicate whether or not an out of limit test is required, and in the former case, the upper and lower bounds of the variable are given. The latest values of the inputs are stored in four specified locations of the scratch pad. The four analogue inputs are processed in two pairs, each associated with a 50 Hz hardware clock interrupt. This corresponds to the maximum operating rate of the multiplexer. For each pair of analogue inputs, the conversion

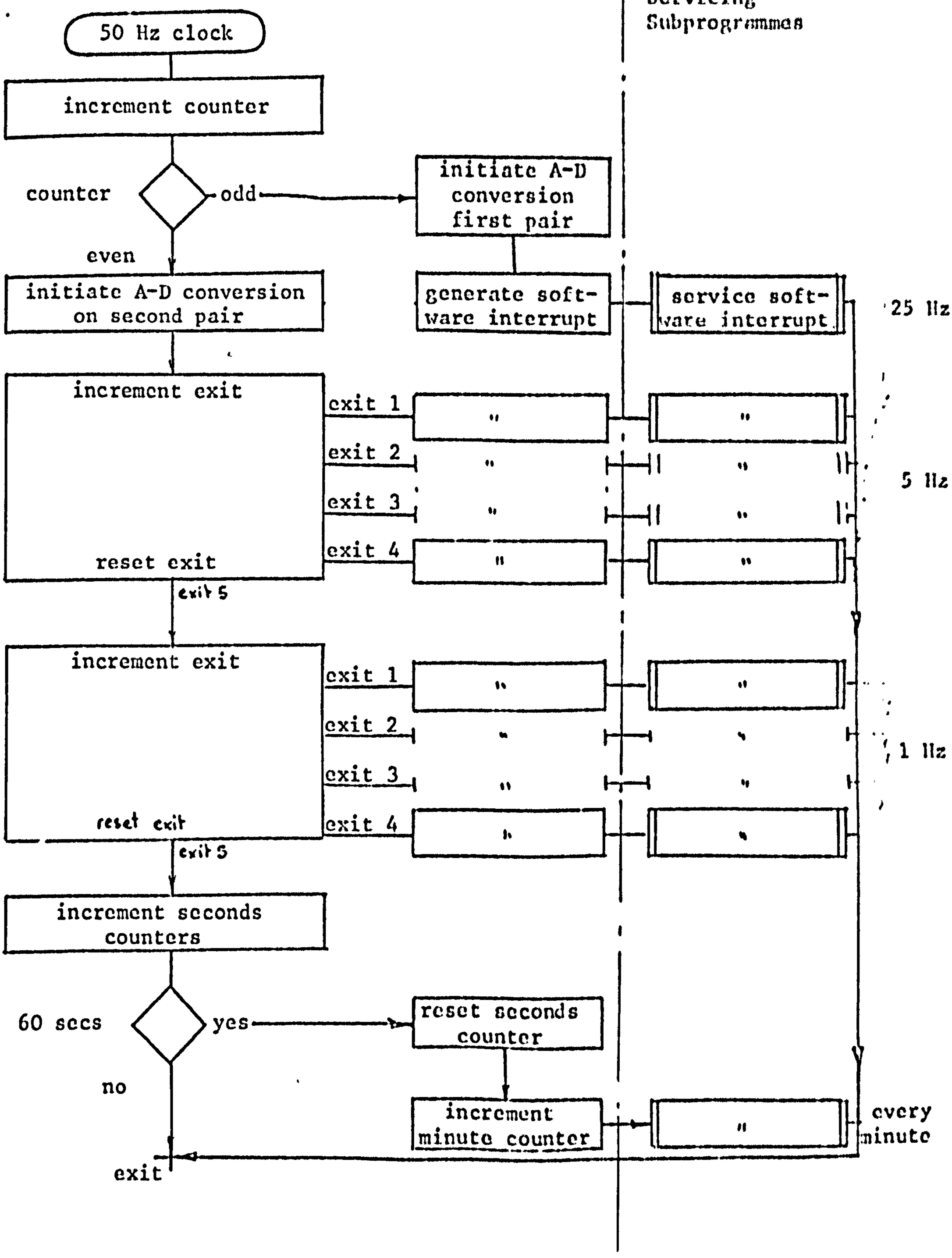


Fig. 4.3.1 - Executive-Timing Section

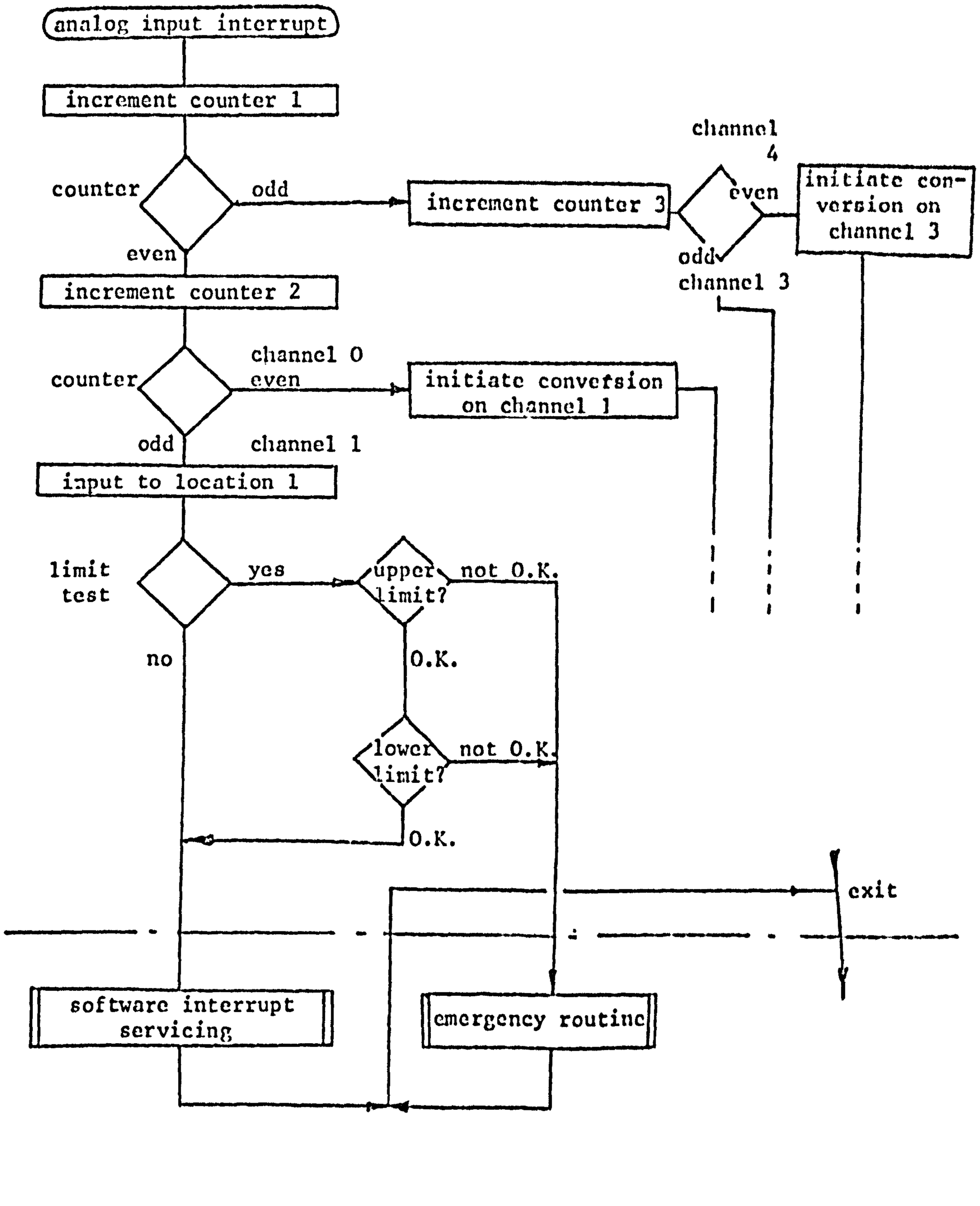


Fig. 4.3.2 - Executive-analogue input interrupt processing

of the first input is initiated by a hardware clock interrupt and the second is initiated by the completion of the first. When a conversion completed interrupt is received, the contents of the converter are transferred to the memory location associated with its position in the input sequence. When the corresponding flag is set, an alarm software interrupt is generated if the variable lies outside the prescribed limits. In all other cases, a software interrupt indication the presence of a new input value is given.

Input by the operator is presented in two forms, directives indicating programme action and data to be stored for future use. No interlace facility is available on the computer. However, a hardware interrupt occurs when the character transfer buffer empties on input and when the peripheral has completed its operation on a word on output. A subprogramme is included in the executive to input and output simple words or strings to and from any typewriter, tape reader or punch utilising this interrupt. Another subprogramme uses the above subprogramme to input a character string terminated by a carriage return, spaces being ignored and an erase causes re-entry to the subprogramme to read a corrected string. The first four characters are then compared with a directory and programme control transferred to the corresponding address when coincidence is found.

A hardware interrupt is reserved for emergency actions and a low priority hardware interrupt is used to allow recovery to the executive from background computation.

4.3.3 Executive Directives

These directives are used for operator or tape control of timing, input, output and other functions of general application. Most directives are followed by parameters specifying further information required for their execution. The directives available are:

TYPE - select the keyboard as the current input peripheral.

TAPE - select the tape reader as the current input peripheral.

READ - read a list into the specified block of store using the current input peripheral.

LIST - output a list from the specified block of store on paper tape.

HOLD - suspend input on the current peripheral. A message *Input suspended* is given on the typewriter. The operator may return control to the current input peripheral by means of the background computation recovery interrupt.

WAIT - suspend input on the current peripheral for the specified number of seconds.

TIME - print time in minutes and seconds since last reset.
If followed by an asterisk, the counter is reset.

CHECK - ensure that the process is operating within bounds. A software interrupt is generated if the single bit digital inputs do not correspond to the specified word or if selected analogue inputs lay outside their defined limits.

The typewriter is made the current input peripheral if such a failure occurs.

AXES - draw axes on the X-Y plotter using two analog outputs to drive the plotter with a relay to raise and lower the pen.

PLOT - plot the specified graph on the X-Y plotter. Constants, incremental constants and lists may all be used as variables and there are three symbols, of adjustable size, available.

ALOG - log the specified channels. Up to four channels may be logged at similar rates. The last of the four analogue input channel address locations and its associated software interrupt are used. Control is transferred to the current input peripheral on completion.

4.3.4 Engine Directives

START - attempt to start engine. Three attempts are made by the computer to start the machine. On starting, the operator is informed and a software interrupt generated. If the engine fails to start, the operator is informed and control transferred to the executive hold directive. The flow diagram for this programme is given in fig. 4.3.3.

STOP - stop the engine. The ignition is turned off and the speed limit check halted. The operator is informed.

IDLE - idle engine. Idling conditions are set up on the engine actuators using the analogue outputs.

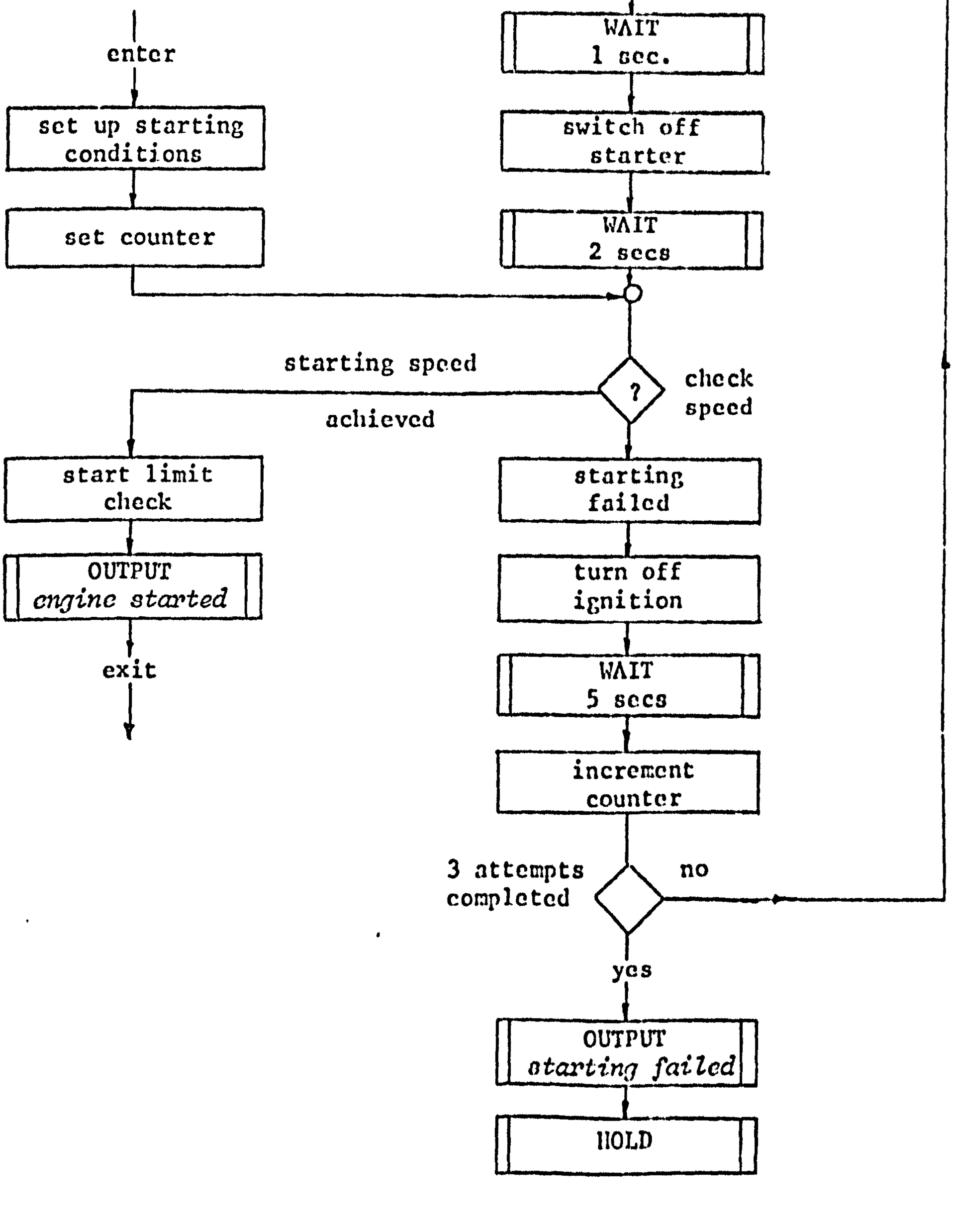


Figure 4.3.3 Flow Diagram for Start Engine Directive

4.3.5 Subprogrammes

The purpose of the subprogrammes is to service the software interrupts generated by the executive and directives. The GOTO directive may be used to enter an initialising routine before the subprogramme becomes active.

The optimisation subprogramme is of particular importance. When supplied with the appropriate subroutines, this outputs a perturbation, correlates the response and uses the correlation results for optimising any specified variable. Others include a stall checking subprogramme, an engine speed control subprogramme, an elementary interrupt clearing subprogramme and a subprogramme for calculating mean and variance of any logged variable.

The optimisation programmes use the simple optimiser described in chapter 2,with the least sequences identification schemes derived in section 2.7 for square waves, p.r.b.s. and p.r.t.s.

CHAPTER 5

Experiments and Results

Introduction

This chapter describes a series of experiments carried out to obtain qualitative confirmation of the previous theoretical analysis. The optimisers, implemented on a digital computer, were used to maximise the power of the internal combustion engine on the test rig by adjusting the ignition angle while the throttle was fully opened. The experiment was completely controlled by the computer and using this configuration, a plot of the optimum ignition setting versus engine speed could be obtained by providing a suitable test tape. This would give a result useful to the designers of ignition timing devices.

The optimiser of chapter 2 is usefully applied, whilst the two derivative hill-climber of chapter 3 is shown to be inapplicable in conjunction with the optimisation of ignition angle.

5.1 General

5.1.1 Power Characteristic

Power is given by the product of speed and torque. If the ignition setting is perturbed, the computation of power from the speed and torque measurements will introduce a non-linearity after the dynamics which will cause errors in the identification of the system dynamics. However, in the steady state the power P is given by

$P = \tau\omega$ for a torque of τ and angular velocity ω,

and hence the gain between the power and the ignition setting θ is given by

$$\frac{\partial P}{\partial \theta} = \omega\frac{\partial \tau}{\partial \theta} + \tau\frac{\partial \omega}{\partial \theta}$$

Providing a good speed controller is employed, the variation of speed with ignition setting should be negligible in the steady state and therefore

$$\frac{\partial P}{\partial \theta} \propto \frac{\partial \tau}{\partial \theta}$$

Then it is only necessary to optimise the torque to achieve maximum power.

5.1.2 Experimental Configuration

For all the experiments, the engine speed was controlled by the dynamometer load with a controller implemented in the digital computer. The throttle setting and desired speed could either be set manually through the keyboard of the on-line computer or via paper tape. The full executive with all directives was available allowing automatic start-up and warm-up periods together with the continued checking of temperatures, speeds and the dynamometer loading necessary for the safe operation of the plant. The additional programmes available were a general optimiser and correlaters, perturbation generators and gain estimators for pseudo-random binary sequences, pseudo-random ternary sequences and square waves. The experiments were all performed with the same engine speed controller set point to allow comparison of the results.

5.2 Preliminary Experiments

5.2.1 Static and Dynamic Characteristics

Experiments were carried out to determine the system characteristics pertinent to the operation of the optimiser. Firstly the static characteristics relating the power output to the ignition angle for constant speed and throttle settings were obtained for a range of engine speeds and throttle settings. Fig. 5.2.1 shows a typical result. In general, as the speed is lowered, the optimum ignition setting becomes less advanced and the hill flattens. Similarly, the effect of increasing the throttle setting is to require less ignition advance for peak power.

Sine wave testing techniques were used to investigate the dynamic characteristics of the rig with the simple computer operated speed controller connected. The results are shown in fig. 5.2.2. The dominant feature is a resonant peak at 9·5 Hz and it can be seen that there is a flat response with this speed controller for frequencies lower than 2 Hz.

5.2.2 Noise Characteristics

Finally the noise characteristics of the torque measurement with the speed controller connected were determined to help evaluation of the results. The torque measurement was monitored whilst the input variables were held constant and 4095 samples of this measurement recorded on paper tape. An off-line analysis was then

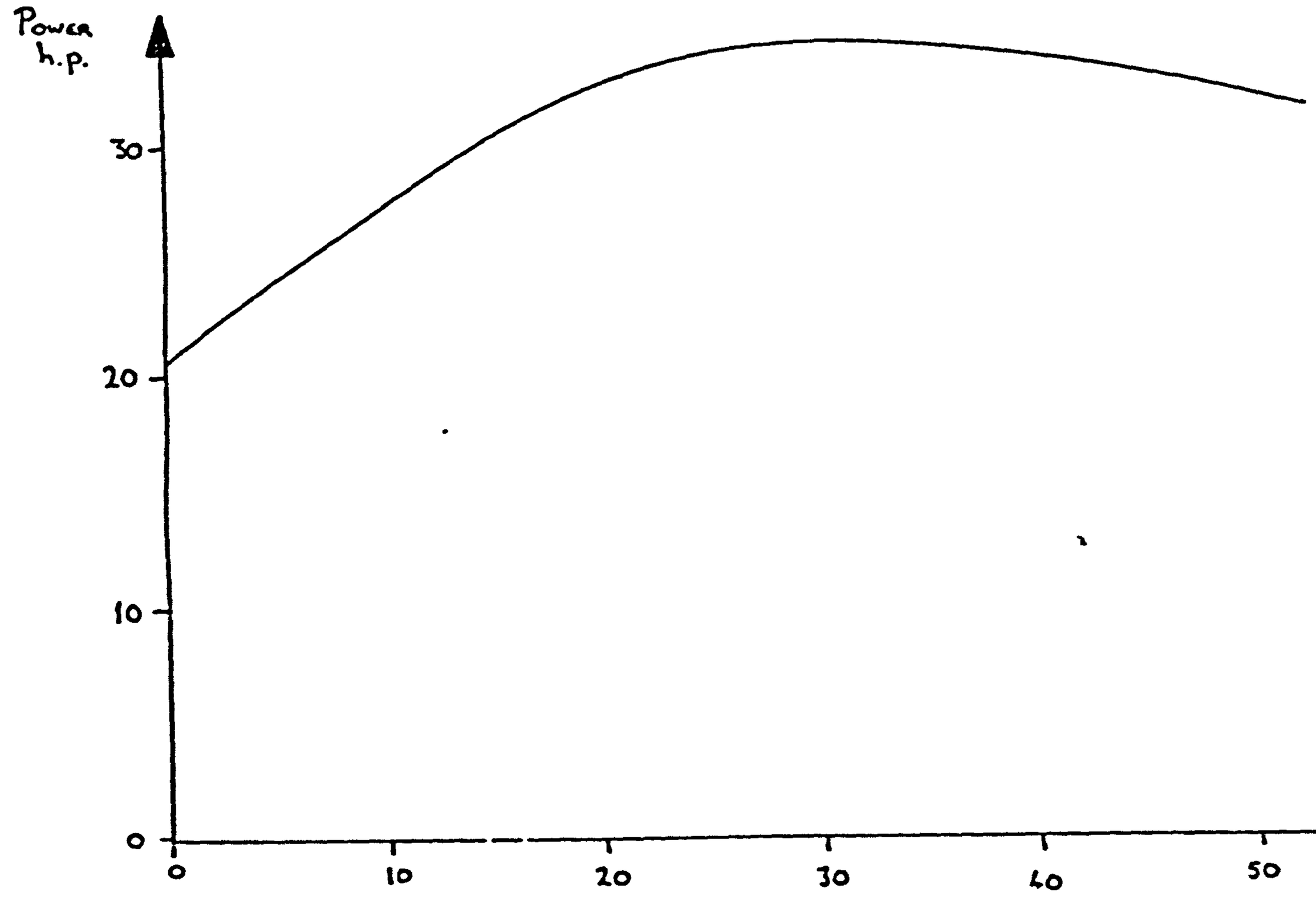
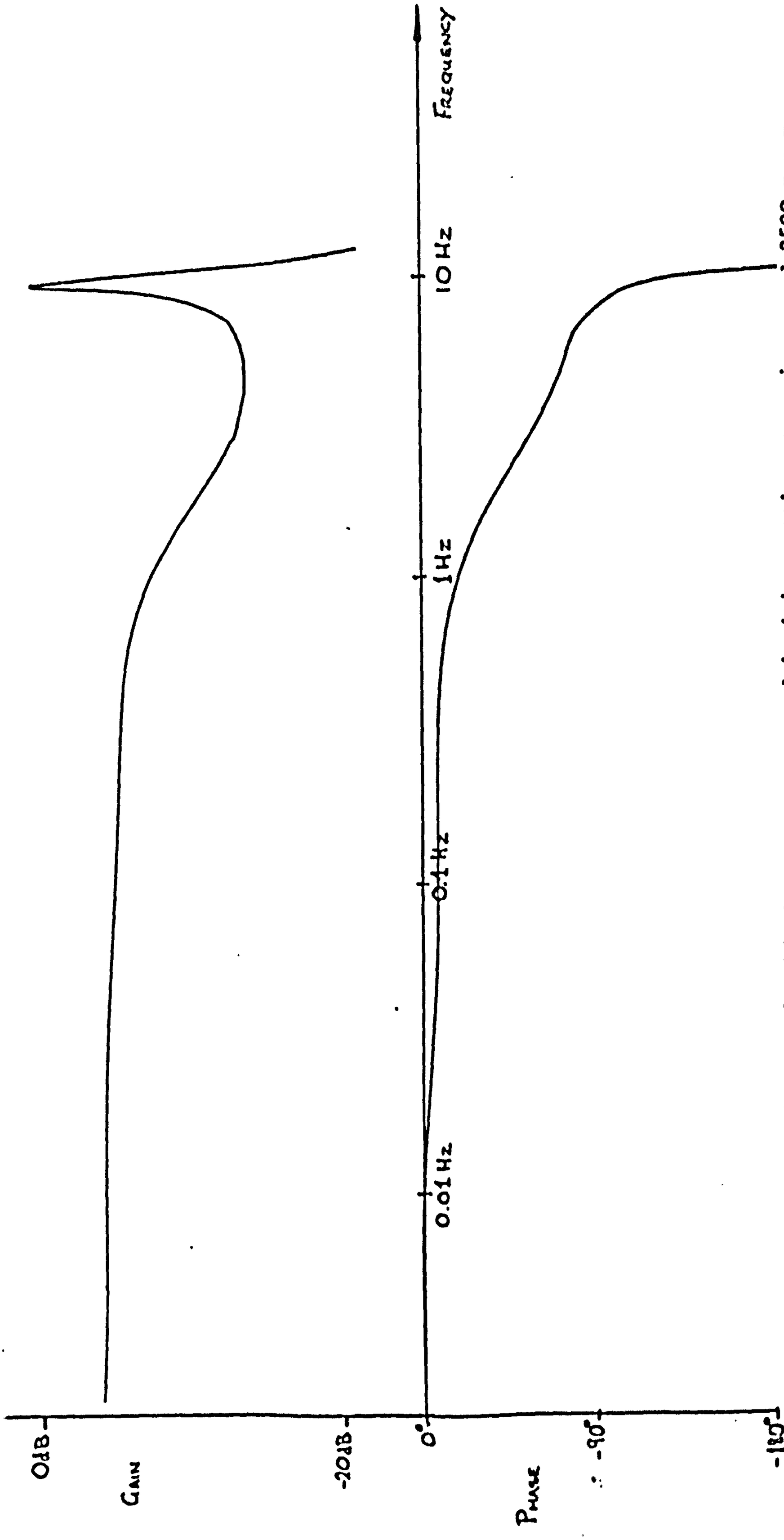


Fig. 5.2.1 Static Relationship between power and ignition setting at engine speed 2500 r.p.m. and full throttle



Dynamic relationship between torque and ignition setting at engine speed 2500 r.p.m.,
Full throttle applied and speed controller connected

Fig. 5.2.2

performed to establish the power spectrum, amplitude probability distribution function and the effect of the ignition setting on the variance of the noise.

The power spectrum shown in fig. 5.2.3 is not that of white noise. It resembles the dynamic characteristics between ignition setting and torque, being reasonably flat below 2 Hz but the resonant peak is not so sharply defined. Some smoothing would, however, be expected as the Fourier transform algorithm used incorporated a smoothing window. The result may be interpreted by assuming that the engine is a source of white noise due to the irregularities in combustion and this is modified by the dynamics of the engine flywheel, shaft and the dynamometer.

The amplitude probability distribution function is shown in fig. 5.2.4. The distribution is skew which may be the true distribution or caused by noise with a symmetrical distribution plus a low frequency drift. The second possibility was eliminated by taking a long sequence of data and determining the probability distributions for sections of this data. A low frequency drift would have produced similar distributions displaced from one another by different mean values, whereas the truly skew distribution gave a series of identical distributions.

As the ignition setting is increased, the variance of the noise was found to tend towards a minimum and then increase. The minimum variance corresponds approximately with the ignition setting that produces maximum power. The falling off in power can be partially explained by the inefficient combustion conditions which cause uneven operation from one combustion to the next. This also

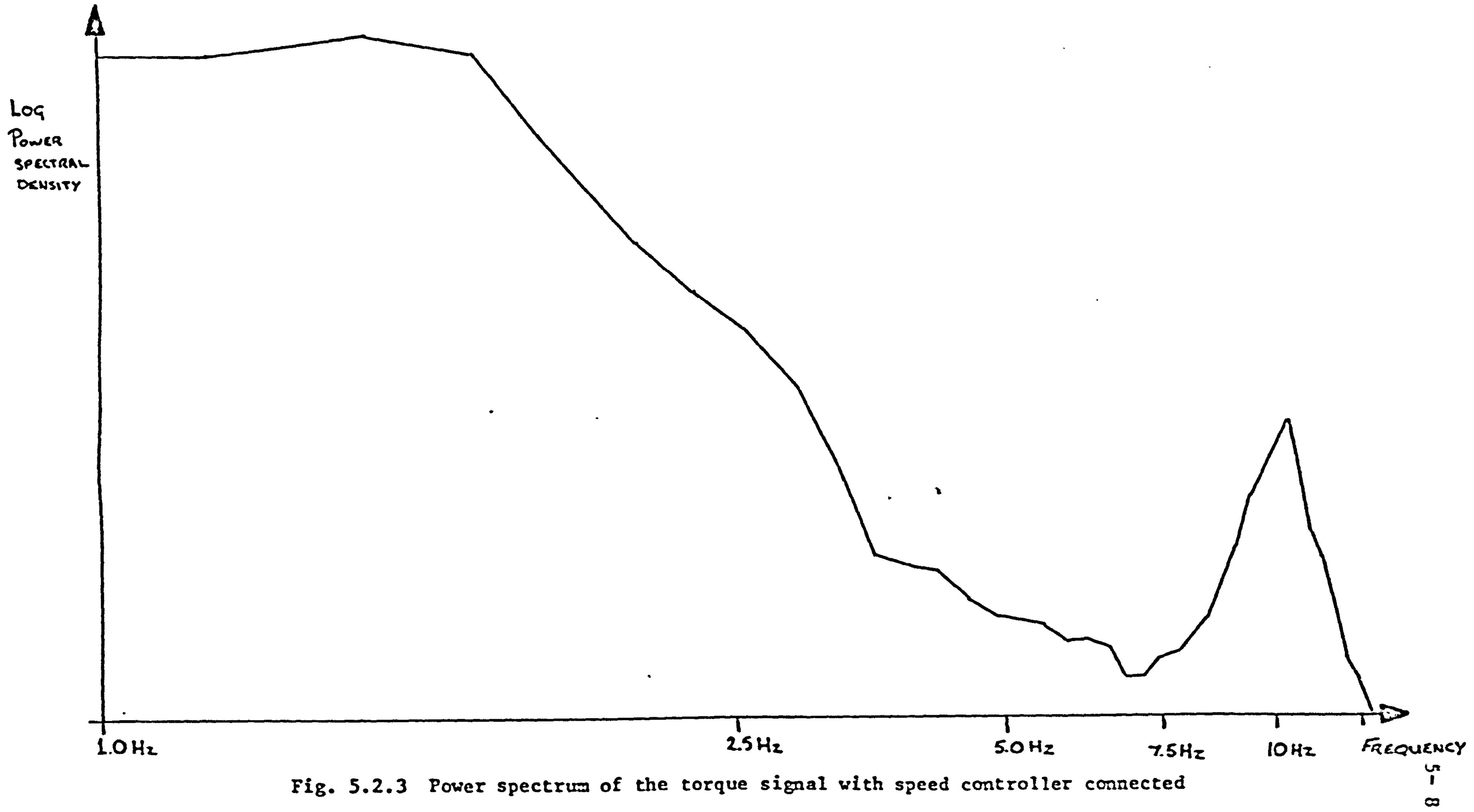


Fig. 5.2.3 Power spectrum of the torque signal with speed controller connected

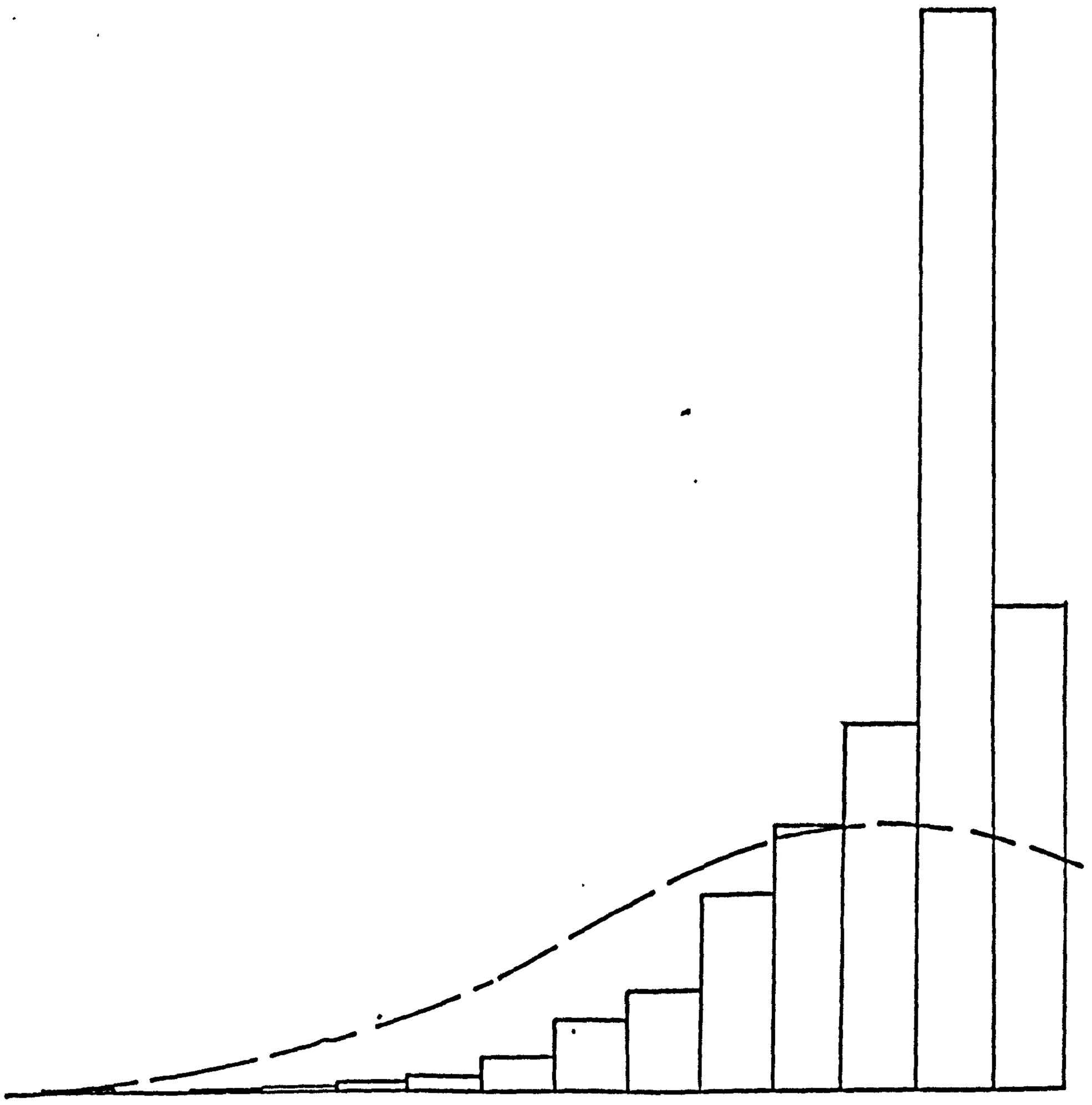


Fig. 5.2.4 Amplitude Probability Distribution Function of the Torque Signal with Speed Controller connected (broken line shows the normal distribution which gives minimum Chi-squared value)

introduces an unevenesss in the torque supplied and hence increases the variance in the torque delivered to the engine. This may become severe at the bounds of ignition setting where the charge occasionally completes ignition in the exhaust system or at the other limit, ignition occurs before the closure of the inlet valve.

5.3 Experimental Work

Pseudo-random binary sequences were used to perturb the ignition setting in order to show that the identification of the ignition angle/torque path was possible. Fig. 5.3.1 shows the correlations obtained with this experiment together with the derived step responses. The experiment was then repeated using 3-level sequences and a similar result obtained. When the response to a 3-level perturbation was correlated with the square of the sequence, however, noise completely masked any result. Further experiments with larger amplitude sequences revealed a constant irregular pattern which could only be modified by using a different 3-level sequence. This suggests that non-linearities were affecting the results and the 2-derivative hill climber previously described could not be implemented.

A series of experiments were carried out using pseudo-random binary sequences, pseudo-random ternary sequences and square waves in the simple optimiser of Chapter 2 and the results of section 2.7. Prior to each experiment, the dynamic and static characteristics were checked after the warm-up period and it was found that the static characteristic varied with time. Measurements showed some correlation between these variations and the ambient temperature and pressure. Humidity was not measured but it seems likely that the variations were due to the effect of changing atmospheric conditions on the combustion and pumping action of the engine.

Experiments were carried out to determine the effects of optimiser gain, perturbation amplitude and averaging the system

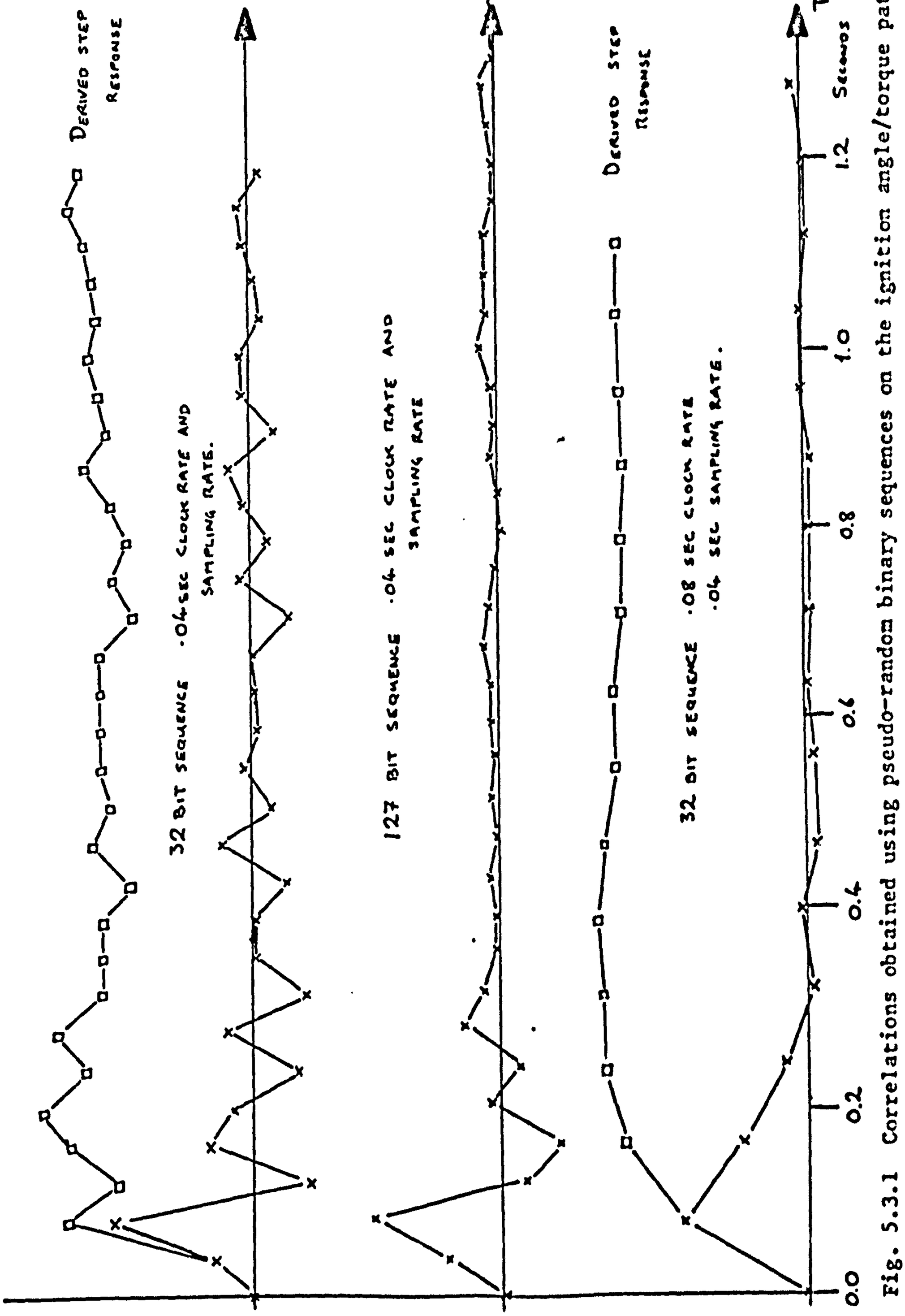


Fig. 5.3.1 Correlations obtained using pseudo-random binary sequences on the ignition angle/torque path

CHAPTER 6

Conclusions

The analysis of the simple optimiser has shown that the value of loop gain has a considerable effect on the dynamic performance of the optimiser and its sensitivity to noise. The maximum value of the first derivative of the cost function is therefore needed for the design of optimisers of this type. The optimisers using higher order models give a more consistent performance but the application of three-level sequences is too severely restricted to particular system configurations to be of great practical use. A study of other methods of implementing higher order model optimisers would be useful.

It has been shown that maximal-length sequences and square waves are suitable for the estimation of system gain. Square waves offer simpler methods of generation and a better performance in the presence of noise, whereas the use of pseudo-random binary sequences gives a more rapid estimate of system gain when noise is of little consequence. The least squares methods used in the analysis were useful for minimising the effects of disturbances and only require a knowledge of the system settling time to design an experiment based on the results obtained.

The practical results tend to confirm the theoretical studies but the occasional changes in the engine characteristics due to atmospheric conditions do not allow a direct comparison of the wave-forms from the experiments carried out. A more consistent performance

could be obtained by controlling the intake air conditions and the exhaust back pressure.

The assumptions about the noise made in the theoretical analysis were not satisfied by the torque signal from the engine test rig. The estimates may be inefficient because the noise is non-white, but this is overcome by using slower clock rates and averaging over several samples. The bias in the probability distribution leads to bias in the gain estimate and ultimately introduces a small error in detecting the optimum. A solution to these problems is to apply maximum likelihood methods, but these require considerable computer time and space and would have to adapt themselves to noise characteristics which differ over the range of operating conditions. The experiments show that the simplifying assumptions about the noise required for least squares still allow acceptable hill climbers.

Acknowledgements

The instrumentation and modelling of the engine test rig were carried out jointly with J. V. Comfort who must be further acknowledged for designing and building the rig and for his help during the experimental work. I would also like to thank Professor J. L. Douce for his guidance throughout this research and Rosemary A. Went for her help in the presentation and typing of the script.

The work presented was made possible by the support of the Ministry of Technology and the facilities provided by the University of Warwick School of Engineering Science.

References

1 Van der GRINTEN, P. M. E. M.

The application of random test signals in process identification

Proc. IFAC congress, Basle, 1963, Butterworths and Co. (Publishers) Ltd.

2 DOUCE, J. L. and NG, K. C.

The use of pseudo-random signals in adaptive control

IFAC Symposium, Teddington, 1965

3 DRAPER, C. S. and LI, Y. T.

Principles of optimalizing control systems and an application to the internal combustion engine

ASME Special publication, Sept 1951

4 ELSPAS, B.

The theory of autonomous linear sequential networks

IRE Trans on circuit theory, Vol. CT6, March 1959, p 45

5 JOHNSTON, J.

Econometric methods

McGraw-Hill Book Company Inc., New York, 1963

6 DAVIES, W. D. T.

Identification of a system in the presence of low frequency drift

Electronics Lts., Vol. 2, 1966, p 327

7 MACLEOD, C. J.

Method of minimising the effect of disturbances on the estimate of the linear impulse response of a linear system

Electronics Lts. Vol. 4, 1968, p 220

8 WILSON, H.

Automatic elimination of bias in binary cross-correlation experiments

Electronics Lts., Vol. 3, 1967, p 514

9 CLARKE, D. W.

Self-optimising systems involving the estimation of cost function slope using pseudo-random binary perturbations

NPL Autonomics div report, June 1966

10 THEILHEIMER, F.

A matrix version of the fast Fourier transform

IEEE Trans. audio and electroacoustics, Vol. AU-17, June 1969, p 158

11 HAZELRIGG, A. P. G. and NOTON, A. R. M.

Application of cross-correlating equipment to linear system identification

Proc. IEE, Vol. 112, 1965, p 2385

12 DOUCE, J. L., NG, K. C. and WALKER, A. E. G.

System identification in the presence of ramp disturbances

Electronics Lts., Vol. 2, 1966, p 243

13 BARKER, H. A.

Choice of pseudo-random binary signals for system identification

Electronics Lts., Vol. 3, 1967, p 524

14 REAM, N.

Proof of the drift resistant property of binary m-sequences

Electronics Lts., Vol. 4, 1968, p 380

15 DAVIES, W. D. T. and DOUCE, J. L.

On-line system identification in the presence of drift

IFAC Prague, 1967, Paper 3.12

16 BROWN, R. F.

Drift correction in periodic cross-correlation schemes

Electronics Lts., Vol. 4, 1968, p 478

17 BROWN, R. F.

Review and comparison of drift correction schemes for periodic cross-correlation

Electronics Lts., Vol. 5, 1969, p 179

18 GARDINER, A. B.

Elimination of the effect of non-linearities on process cross-correlations

Electronics Lts., Vol. 2, 1966, p 164

19 BRIGGS, P. A. N. and GODFREY, K. R.

Pseudo-random signals for the dynamic analysis of multivariable systems

Proc. IEE, Vol. 113, No. 7, 1966, p 1259

20 ELSDEN, C. S. and LEY, A. J.

A digital transfer function analyser based on pulse rate techniques

Automatica, Vol. 5, No. 1, Jan 1969

21 AYRES, R.

Matrices

Schaum Publishing Company, New York, 1962, p 56

22 DAVIES, E. J.

An experimental and theoretical study of eddy current couplings and brakes

IEEE Trans. Power Apparatus Syst., 1963, p 401

23 DAVIES, E. J.

General theory of eddy current couplings and brakes

Proc IEE, Vol. 113, No. 5, 1966, p 825

24 GIBBS

Theory and design of eddy current slip couplings

Beama. J., 1946, p 123, 172, 219

25 LINDORFF, D.

Theory of sampled-data control systems

John Wiley and Sons, New York, 1965

26 CLARKE, D. W. and Godfrey, K. R.

Simultaneous estimation of the first and second derivatives of a cost function

Electronics Ltd., Vol. 2, 1966, p 338

27 CLARKE, D. W. and GODFREY, K. R.

Three level m-sequences and their application in on-line hill climbing

IEE Colloquium on Pseudo-random sequences, 1967, p 1-1

28 ROBERTS, J. D.

Extremum on hill climbing regulation:- a statistical theory involving lags, disturbances and noise

Proc. IEE, Vol. 112, No. 1, Jan 1965, p 137

29 DOUCE, J. L. and KING, R. E.

A self-optimising non-linear control system

Proc. IEE, Vol. 108B, July 1961, p 441

30 DOUCE, J. L. and BOND, A. D.

The development and performance of a self-optimising system

Proc. IEE, Vol. 110, No. 3, 1963, p 619

31 FLETCHER, R. and POWELL, M. J. D.

A rapidly convergent descent method for minimisation

The Computer Journal, Vol. 6, 1963, p 163

32 POWELL, M. J. D.

An efficient method for finding the minimum of a function of several variables without calculating derivatives

UKAEA Research group report AERE-TP138, 1964

33 Van der GRINTEN, P. M. E. M.

Multivariable optimising control by an analogue computer

Trans. Instn. Chem. Engnrs., Vol. 40, 1962, p 356

Symbolic Notation

z	z-transform operator
$\langle u \rangle$	sequence u
u_m	mth element of u
var	variance
covar	covariance
plim	probability limit
E	expected value
$\hat{a}$	least squares estimate of a
$\tilde{a}$	maximum likelihood estimate of a
$\underline{A}$	matrix
$\underline{a}$	column vector
$\underline{A}'$	transpose of $\underline{A}$
$\underline{A}^{-1}$	inverse of $\underline{A}$
$\underline{O}_k$	null matrix
$\underline{I}_k$	unit matrix
$\underline{J}_k$	square matrix with every element unity
$\underline{1}_k$	column vector with every element unity

The dimension suffix k is only used when the dimensions are unclear.

$\oplus_N$	modulo N addition operator
$\ominus_N$	modulo N subtraction operator

APPENDIX A1 - MATRIX INVERSION

A1. 1 - Inversion of $aI + bJ$ matrix

Let the inverse of the matrix $(aI_k + bJ_k)$ be $(cI_k + dJ_k)$

Then $(aI_k + bJ_k)(cI_k + dJ_k) = I_k$

$$acI_k + (bc + ad)J_k + bdJ_k^2 = I_k$$

Substituting $J_k^2 = kJ_k$ and solving, $c = \frac{1}{a}$ and $d = -\frac{b}{a(a + bk)}$

Hence $\left(aI_k + bJ_k\right)^{-1} = \frac{1}{a}\left(I_k - \frac{b}{(a + bk)} J_k\right)$ ... (A1. 1)

A1. 2 - Inversion by partitioning

It can be shown[21] that the inverse of a square matrix A can be found by partitioning.

$$\text{Let } A^{-1} = \left(\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right)^{-1} = B = \left(\begin{array}{c|c} B_{11} & B_{12} \\ \hline B_{21} & B_{22} \end{array}\right)$$

where A_{11} and A_{22} are square.

Then $B_{11} = A_{11}^{-1} + A_{11}^{-1}A_{12}\zeta^{-1}A_{21}A_{11}^{-1}$ ... (A1. 2. 1)

$$B_{12} = -A_{11}^{-1}A_{12}\zeta^{-1} \quad \text{... (A1. 2. 2)}$$

$$B_{21} = -\zeta^{-1}A_{21}A_{11}^{-1} \quad \text{... (A1. 2. 3)}$$

$$B_{22} = \zeta^{-1} \quad \text{... (A1. 2. 4)}$$

where $\zeta = A_{22} - A_{21}A_{11}^{-1}A_{12}$ ... (A1. 2. 5)

A1. 3 - Inversion of a General Form

Consider the matrix $\left(\begin{array}{c|c} a\underline{I}_k + b\underline{J}_k & c\underline{1}_k \\ \hline c\underline{1}_k' & d \end{array}\right)$

whose inverse is given by $\left(\begin{array}{c|c} \underline{D}_{11} & \underline{D}_{12} \\ \hline \underline{D}_{21} & \underline{D}_{22} \end{array}\right)$

Substituting in equation (*A1.2.5*)

$$\underline{\zeta} = d - c\underline{1}_k' \left(a\underline{I}_k + b\underline{J}_k\right)^{-1} \underline{1}_k c$$

Using equation (*A1.1*)

$$\underline{\zeta} = d - \frac{c^2 k}{(a + bk)}$$

Hence, defining $\gamma = bd - c^2$ and substituting in equations (*A1.2.1*) to (*A1.2.4*)

$$\underline{D}_{11} = \frac{1}{a}\left[\underline{I}_k - \frac{\gamma}{(ad + \gamma k)}\underline{J}_k\right] \qquad \ldots (A1.3.1)$$

$$\underline{D}_{12} = -\frac{c}{(ad + \gamma k)}\underline{1}_k \qquad \ldots (A1.3.2)$$

$$\underline{D}_{21} = \underline{D}_{12}' \qquad \ldots (A1.3.3)$$

$$\underline{D}_{22} = \frac{a + bk}{(ad + \gamma k)} \qquad \ldots (A1.3.4)$$

A1. 4 - Inversion of a Second General Form

Let the $(k + 1)$ square matrix $\underline{A}$ be defined by:

$$\underline{A} = \begin{pmatrix} n & n-1 & n-2 & \cdot & \cdot & \cdot & \cdot & n-k \\ n-1 & n & n-1 & & & & & \cdot \\ n-2 & n-1 & n & & & & & \cdot \\ \cdot & & & & & & & \\ \cdot & & & & & & & \\ \cdot & & & & & & n & n-1 \\ n-k & n-k-1 & & & & & n-1 & n \end{pmatrix}$$

The inverse of $\underline{A}$ may be found using the relationship,

$$\underline{A}^{-1} = \underline{B}\,\underline{B}^{-1}\underline{A}^{-1}\left(\underline{B}'\right)^{-1}\underline{B}'$$
$$= \underline{B}\left[\underline{B}'\underline{A}\,\underline{B}\right]^{-1}\underline{B}',$$

where $\underline{B}$ is a suitable, non-singular $(k + 1)$ square matrix.

In this case,

$$\underline{B} = \begin{pmatrix} 1 & 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ -1 & 1 & & & & & \vdots \\ 0 & -1 & & & & & \cdot \\ \cdot & 0 & & & & & \cdot \\ \cdot & \cdot & & & & & \cdot \\ \cdot & \cdot & & & -1 & 1 & 0 \\ 0 & 0 & & & 0 & -1 & 1 \end{pmatrix}$$

Now,

$$\underline{B}'\underline{A} = \begin{pmatrix} 1 & -1 & -1 & \cdot & \cdot & \cdot & \cdot & -1 & -1 \\ 1 & 1 & -1 & & & & & -1 & -1 \\ 1 & 1 & 1 & -1 & & & & -1 & -1 \\ \cdot & & & & & & & \cdot & \cdot \\ \cdot & & & & & & & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & & & 1 & -1 & \cdot \\ 1 & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 & -1 \\ n-k & \cdot & \cdot & \cdot & \cdot & \cdot & n-2 & n-1 & n \end{pmatrix}$$

and

$$\underline{B}'\underline{A}\,\underline{B} = \left(\begin{array}{cccccccccc|c} 2 & 0 & . & . & . & . & . & . & 0 & -1 \\ 0 & 2 & & & & & & & . & . \\ . & & & & & & & & . & . \\ . & & & & & & & & . & . \\ . & & & & & & & & . & . \\ . & & & & & & & 2 & 0 & . \\ 0 & . & . & . & . & . & . & 0 & 2 & -1 \\ \hline -1 & . & . & . & . & . & . & & -1 & n \end{array}\right).$$

Substituting in equations (A1.3.1) to (A1.3.4), with $a = 2$, $b = 0$, $c = -1$ and $d = n$, gives

$$\left[\underline{B}'\underline{A}\,\underline{B}\right]^{-1} = \frac{1}{(2n-k)} \left(\begin{array}{ccccc|c} & & & & & 1 \\ & & & & & . \\ & \tfrac{1}{2}\left((2n-k)\underline{I}_k + \underline{J}_k\right) & & & & . \\ & & & & & . \\ & & & & & 1 \\ \hline 1 & . & . & . & 1 & 2 \end{array}\right) = \underline{C} \text{ say.}$$

Thus,

$$\underline{B}\,\underline{C} = \frac{1}{2(2n-k)} \left(\begin{array}{c|ccccc|c} 2n-k+1 & 1 & 1 & & & 1 & 2 \\ \hline -(2n-k) & (2n-k) & 0 & . & . & 0 & 0 \\ 0 & -(2n-k) & (2n-k) & & & & \\ . & 0 & -(2n-k) & & & & \\ . & . & 0 & & & & \\ . & . & . & . & (2n-k) & & \\ 0 & 0 & 0 & & -(2n-k) & (2n-k) & \\ \hline 1 & 1 & 1 & & 1 & -(2n-k-1) & 2 \end{array}\right)$$

and, hence,

$$\underline{A}^{-1} = \underline{B}\,\underline{C}\,\underline{B}' =$$

$$\frac{1}{2(2n-k)}\begin{bmatrix} 2n-k+1 & -(2n-k) & 0 & \cdots & 0 & 1 \\ -(2n-k) & 2(2n-k) & -(2n-k) & 0 & 0 & 0 \\ 0 & -(2n-k) & 2(2n-k) & & & \\ \cdot & 0 & -(2n-k) & & & \\ \cdot & \cdot & 0 & & & \\ \cdot & \cdot & \cdot & & 2(2n-k) & -(2n-k) \\ 0 & 0 & 0 & 0 & -(2n-k) & 2(2n-k) \\ 1 & 0 & 0 & 0 & -2n-k & 2n-k+1 \end{bmatrix}$$

... (A1. 4. 1)

Appendix 2 Detailed Operation of Instruments

A2.1 Ignition timing unit

A photoelectric transducer and an annular plate mounted on the flywheel monitored the angular movement of the crankshaft. The outer rim of the annulus passed through the transducer and was drilled in two pitch circles, the outer having holes spaced at 1^{o} intervals and the inner, two holes 180^{o} apart. The transducer and flywheel plate are shown in fig. A2.1.1. The square wave generated by the outer transducer had a period corresponding to 1^{o} of crankshaft rotation and was passed through an *and* gate to a six bit counter. Pulses from the inner transducer, occurring at 50^{o} before top dead centre, set a bistable which opened the *and* gate to initiate counting. When the contents of the counter exceeded the desired value, either in analogue or digital form, a monostable generated a pulse to fire the ignition and to reset the bistable to prevent further pulses entering the counter. A further monostable reset the counter for the next cycle of operation.

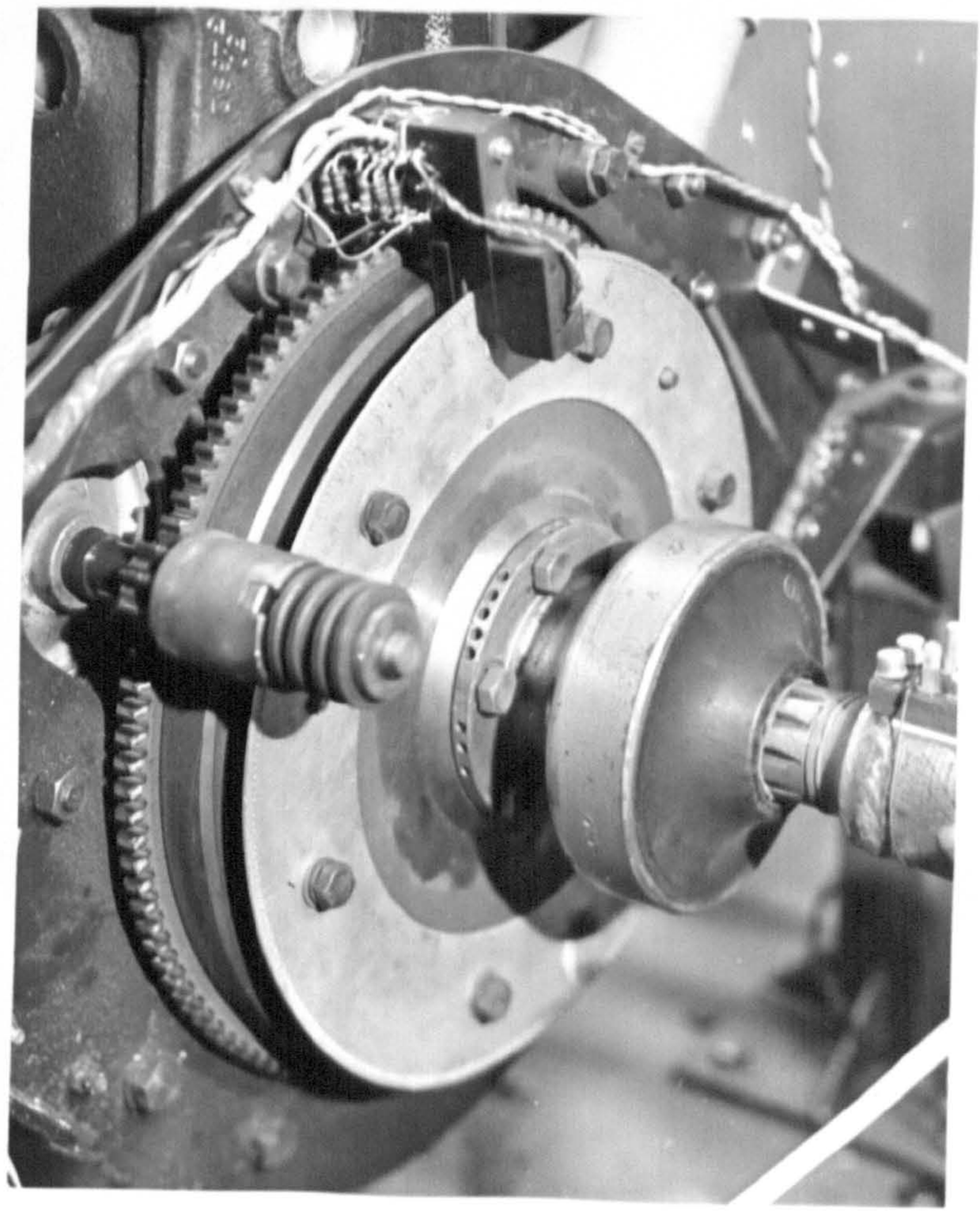


Fig. A2.1.1 Photoelectric Transducer and Flywheel Plate

A2.2 Torque Transducer

With reference to fig. A2.2.1, the difference voltage from the bridge was integrated by the differential integrator until the output of the integrator caused the comparator to switch. The polarity of the supply to the bridge and the comparator level were then reversed. The difference voltage from the bridge was then of opposite polarity and integration continued in the opposite direction until the comparator level was again reached and the process repeated. Any changes in the out of balance of the bridge resulted in a steeper ramp at the output of the integrator causing the circuit to operate at a higher frequency.

The resistance of a strain gauge under tension is given by

$$(1 + \alpha\theta)(1 + \beta\varepsilon)R ,$$

where R is the resistance of the gauge at zero temperature and strain,

α is the temperature coefficient of resistance,

θ is the temperature,

β is the gauge factor

and ε is the strain in the gauge.

The resistances for the bridge configuration shown in fig. A2.2.2 are given for the shaft transmitting torque and a voltage V applied to the bridge. The integrator has input resistors P, capacitors C and output voltage V_0. The two arms of the bridge then have the equivalent circuits shown in fig. A2.2.3 and the output of the differential integrator is therefore given by:

$$V_0 = \frac{1}{C(P + (1 + \alpha\theta)\frac{R}{2})} \int \left[\frac{(1 + \beta\varepsilon)}{2} - \frac{(1 - \beta\varepsilon)}{2}\right] V dt.$$

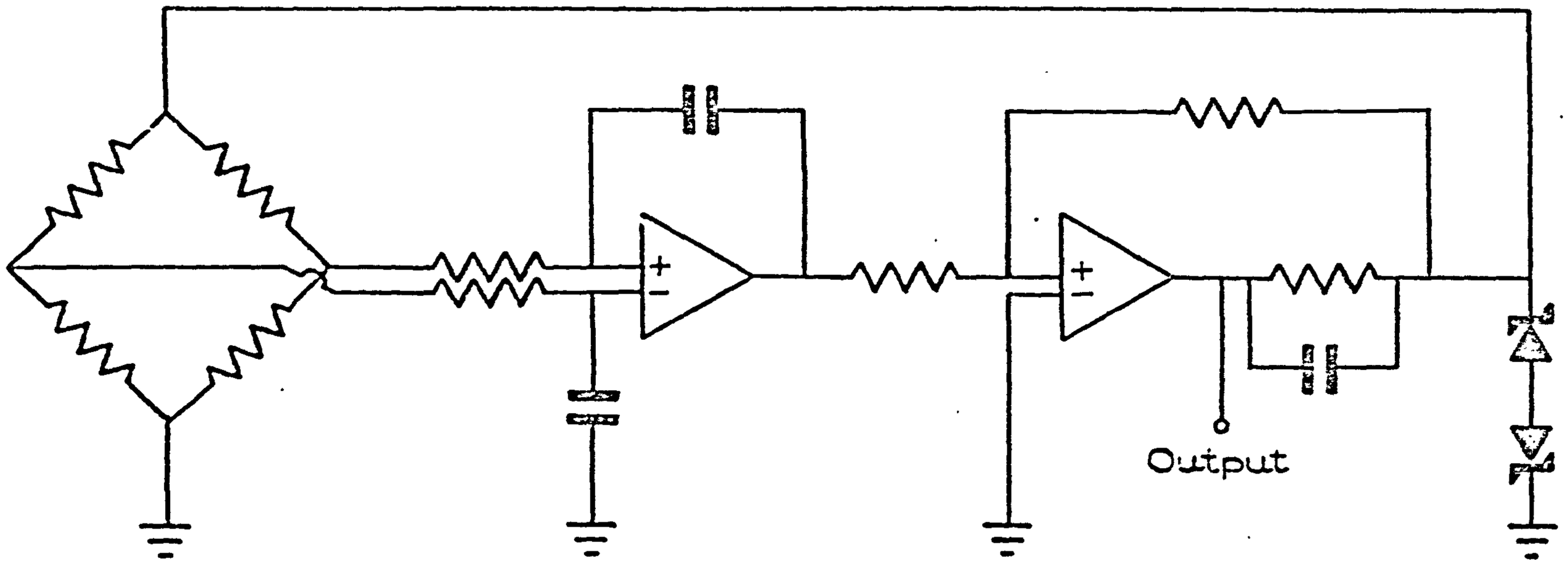


Fig. A.2.2.1 TORQUE TRANSDUCER CIRCUIT DIAGRAM.

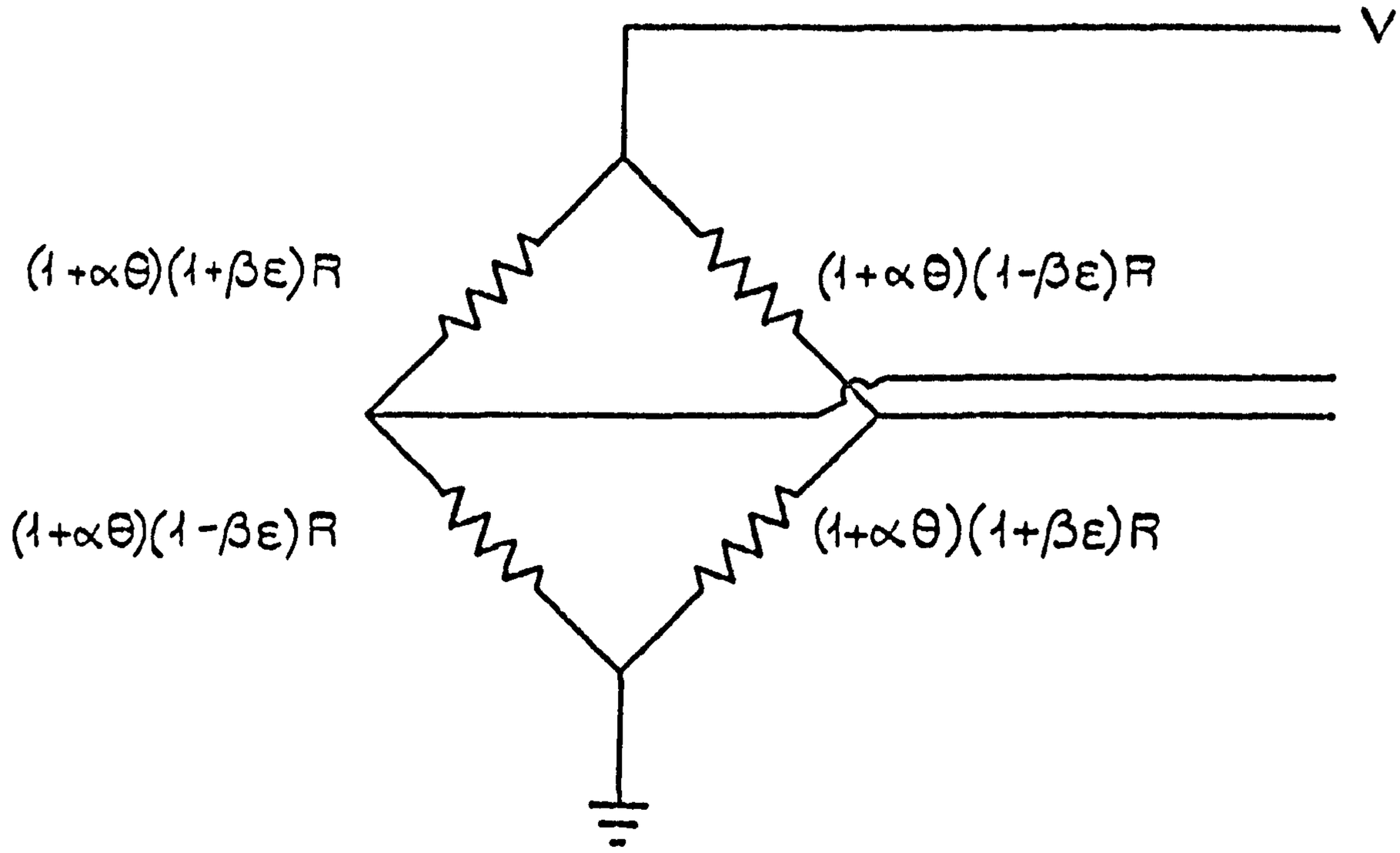


Fig. A.2.2.2 EFFECT OF STRAIN ON BRIDGE RESISTANCE.

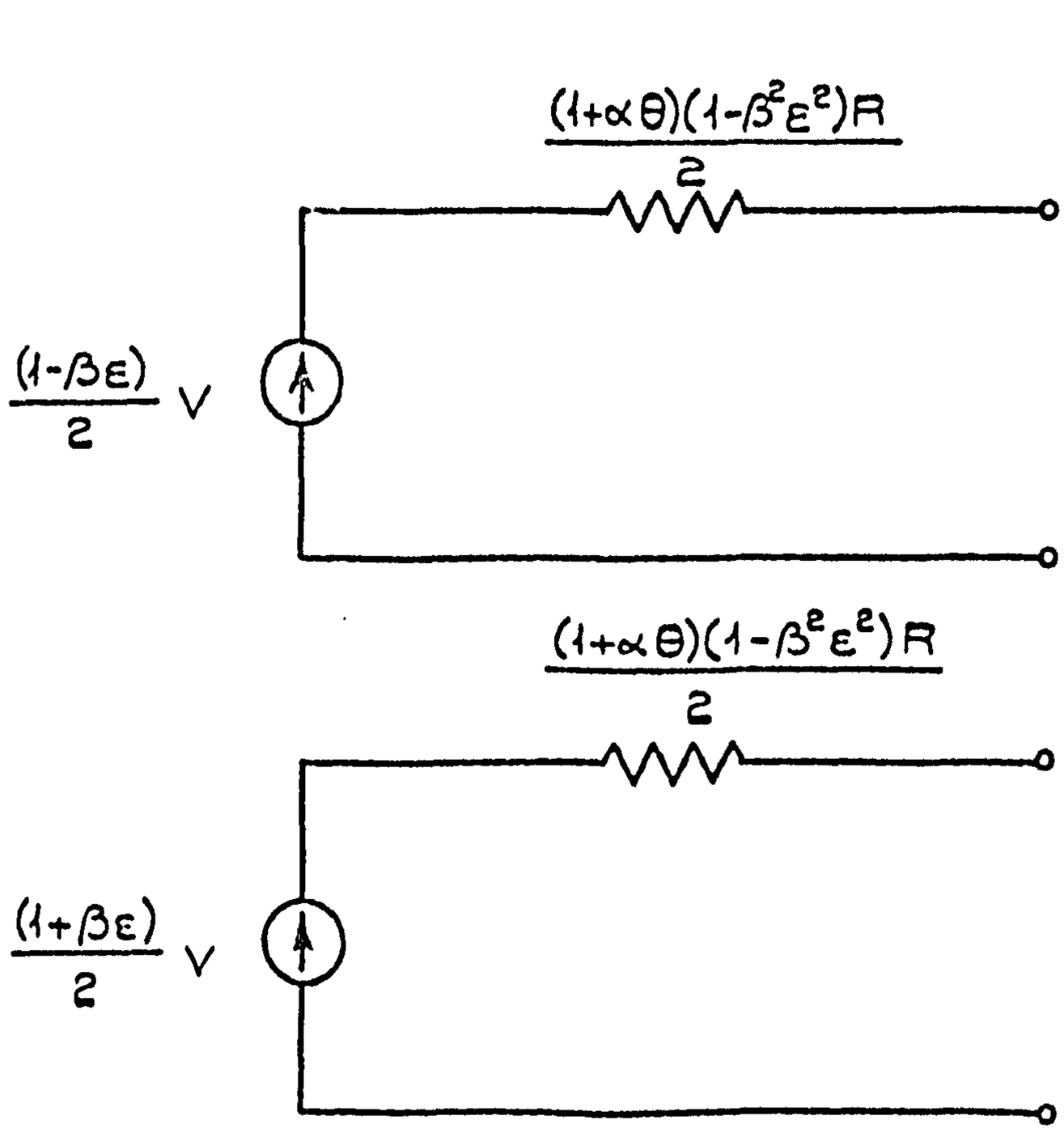


Fig. A.2.2.3 EQUIVALENT CIRCUITS FOR BRIDGE ARMS.

For the strain levels used, $\varepsilon^2\beta^2$ was very small and therefore

$$V_0 = \frac{\beta\varepsilon Vt}{C\left(P + (1 + \alpha 0)\frac{R}{2}\right)} + K,$$

where K is the initial output voltage.

If P is made large compared with R, the output becomes a ramp with gradient $\beta\varepsilon V/CP$. The comparator will trigger at a voltage $V.q$, where q is the ratio of the input to the positive feedback resistance, and therefore the time taken between transitions of the comparator is $2CPq/\beta\varepsilon$ and the output frequency is $\beta\varepsilon/4qCP$. It will be noted that this is independent both of the voltage applied to the bridge and the temperature coefficient of resistance of the strain gauge. The circuit will also compensate for offset currents in the integrator.

A2.3 Precision Monostable

The following conditions existed on the monostable (fig. A2.3.1) in its quiescent state. The positive input to the amplifier was about 0·5*V*. positive due to the current flowing through the reversed Zener diode Z_2, the negative input was at earth potential and the output was therefore positive. When a positive going edge was applied to the input, it was differentiated by the resistor R_1, capacitor C_1 and diode D_1 network and the resulting spike caused the negative input to exceed the positive input of the amplifier. The output voltage then began to decrease, accelerated by the positive feedback through the feedback capacitor C_2, until the Zener diode Z_2 conducted in the usual way several volts negative. The feedback capacitor C_2 then charged at a rate dependent on the resistor R_3, towards the voltage determined by the Zener diode Z_1. Since the Zener diode Z_2 determined the voltage across the capacitor C_2 and the Zener diode Z_1, its charge rate, the timing was made substantially independent of temperature by choosing these Zener diodes to have the same temperature coefficient. When, as a result of the charging of the capacitor C_2, the positive input passed through zero volts, the amplifier switched back to its original state, again accelerated by positive feedback through C_2. The Zener diode Z_2 then represented a very low resistance and the capacitance C_2 discharges rapidly, restoring the circuit for further triggering.

It was essential for accurate timing that negligible current flowed from the operational amplifier while the feedback capacitor C_2 was charging and this was achieved by using an NI'N transistor balanced pair at the input to the operational amplifier. Negative

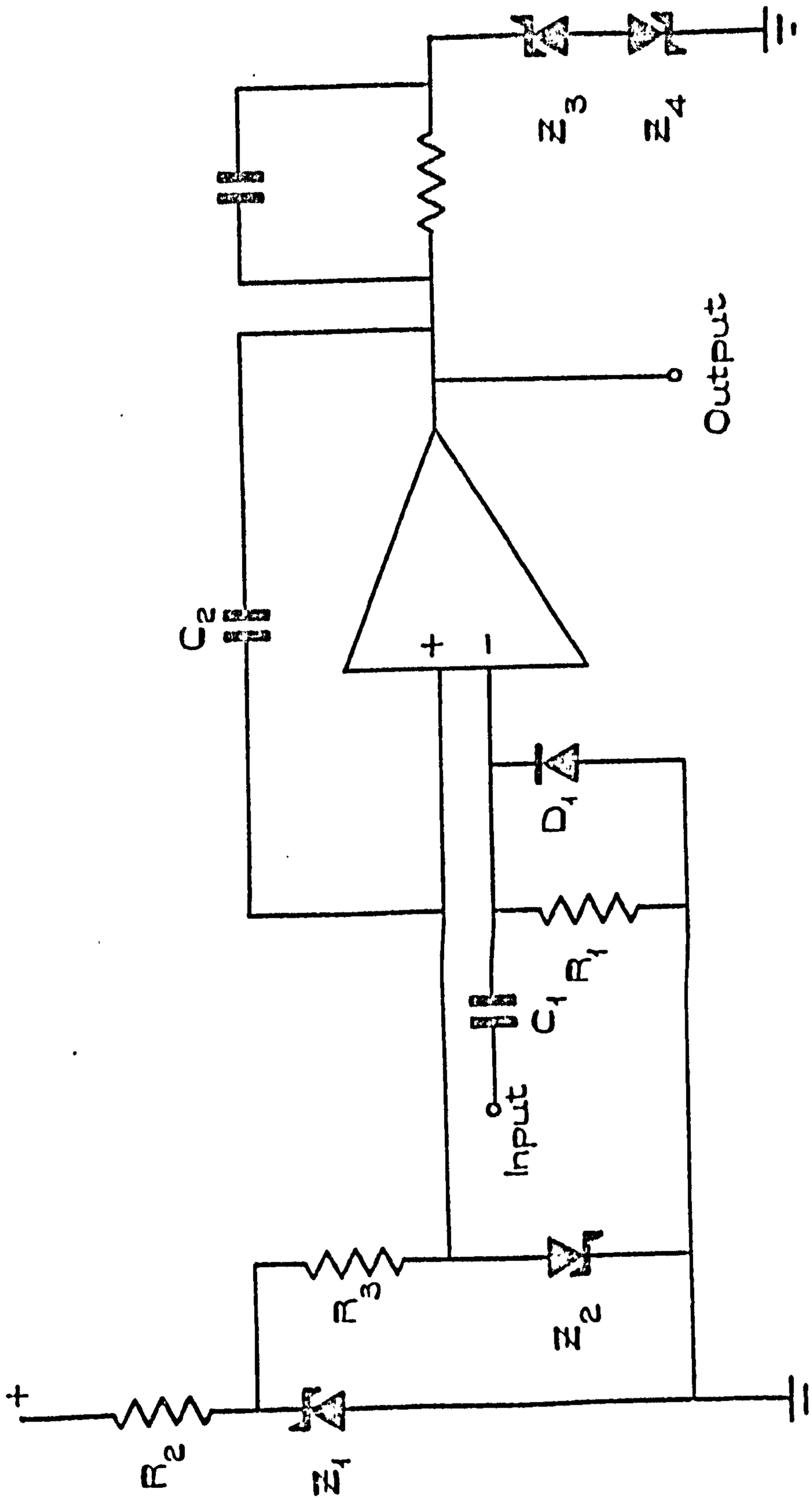


Fig. A.2.3:1 High Stability Monostable.

edges at the input to the monastable were heavily attenuated by the diode D_1 and this permitted the operation of the monostable over a wide range of mark space ratios. The Zener diodes Z_3 and Z_4 were used to stabilise the amplitude of the output.

Appendix A3 Fourth Order Auto-correlation Function of a Three-level Maximal-Length Sequence

The fourth order autocorrelation function of a three-level maximal-length sequence, I_4, is given by

$$I_4 = \frac{1}{T}\int_0^{\infty} u(t - t_1)u(t - t_2)u(t - t_3)u(t - t_4)dt$$

Now $u(T_1)u(T_2) = \left(u(T_1) \oplus_3 u(T_2)\right)^2 \ominus_3 \left(u(T_1) \ominus_3 u(T_2)\right)^2$

and since

$$u(T_1) \ominus_3 u(T_2) = 0, \quad u(T_1) = u(T_2)$$

$$u(T_1) \oplus_3 u(T_2) = 0, \quad u(T_1) = -u(T_2)$$

$$u(T_1)u(T_2) = u^2(T_1) \ominus_3 u^2(T_1), \quad u(T_2) = 0$$

$$\equiv u(T_1) - u(T_1)$$

therefore

$$u(T_1)u(T_2) = \left(u(T_1) \oplus_3 u(T_2)\right)^2 - \left(u(T_1) \ominus_3 u(T_2)\right)^2$$

and the fourth order auto-correlation function becomes

$$I_4 = \frac{1}{T}\int_0^{T} \left\{\left(u(t - t_1) \oplus_3 u(t - t_2)\right)^2 - \left(u(t - t_1) \ominus_3 u(t - t_2)\right)^2\right\}$$
$$\left\{\left(u(t - t_3) \oplus_3 u(t - t_4)\right)^2 - \left(u(t - t_3) \ominus_3 u(t - t_4)\right)^2\right\}dt$$

Using the shift and add property of a three-level maximal-length sequence,

$$u(t - t_1) \oplus_3 u(t - t_2) = u(t - \tau_1)$$

$$u(t - t_3) \oplus_3 u(t - t_4) = u(t - \tau_2)$$

$$u(t - t_1) \ominus_3 u(t - t_2) = u(t - \tau_3)$$

$$u(t - t_3) \ominus_3 u(t \quad t_4) = u(t - \tau_4)$$

giving the fourth order auto-correlation function in the form,

$$I_4 = \phi_{u2u2}(\tau_1 - \tau_2) - \phi_{u2u2}(\tau_1 - \tau_4)$$

$$- \phi_{u2u2}(\tau_3 - \tau_4) + \phi_{u2u2}(\tau_3 - \tau_2)$$

Appendix A4 Detailed Results

A4.1 Pseudo-random binary sequences

A pseudo-random binary sequence of length thirty-one was used throughout these experiments and this was generated by applying feedback from the second and fifth stages of a programmed shift register. The engine speed was controlled using a set point of 2500 r.p.m.

Fig. A4.1.1 Effect of optimiser gain

This group of experiments shows how the optimiser *wander* increases with loop gain.

Perturbation clock rate ·04 sec. Perturbation amplitude ± 2^{o}

The impulse response was assumed to have settled in 15 clock intervals.

Fig. A4.1.2 Effect of perturbation amplitude

These three experiments indicate how the variance of the estimate of the cost function slope increases with smaller perturbation amplitude.

Perturbation clock rate ·04 sec. Gain factor 2

The impulse response was assumed to have settled in 15 clock intervals.

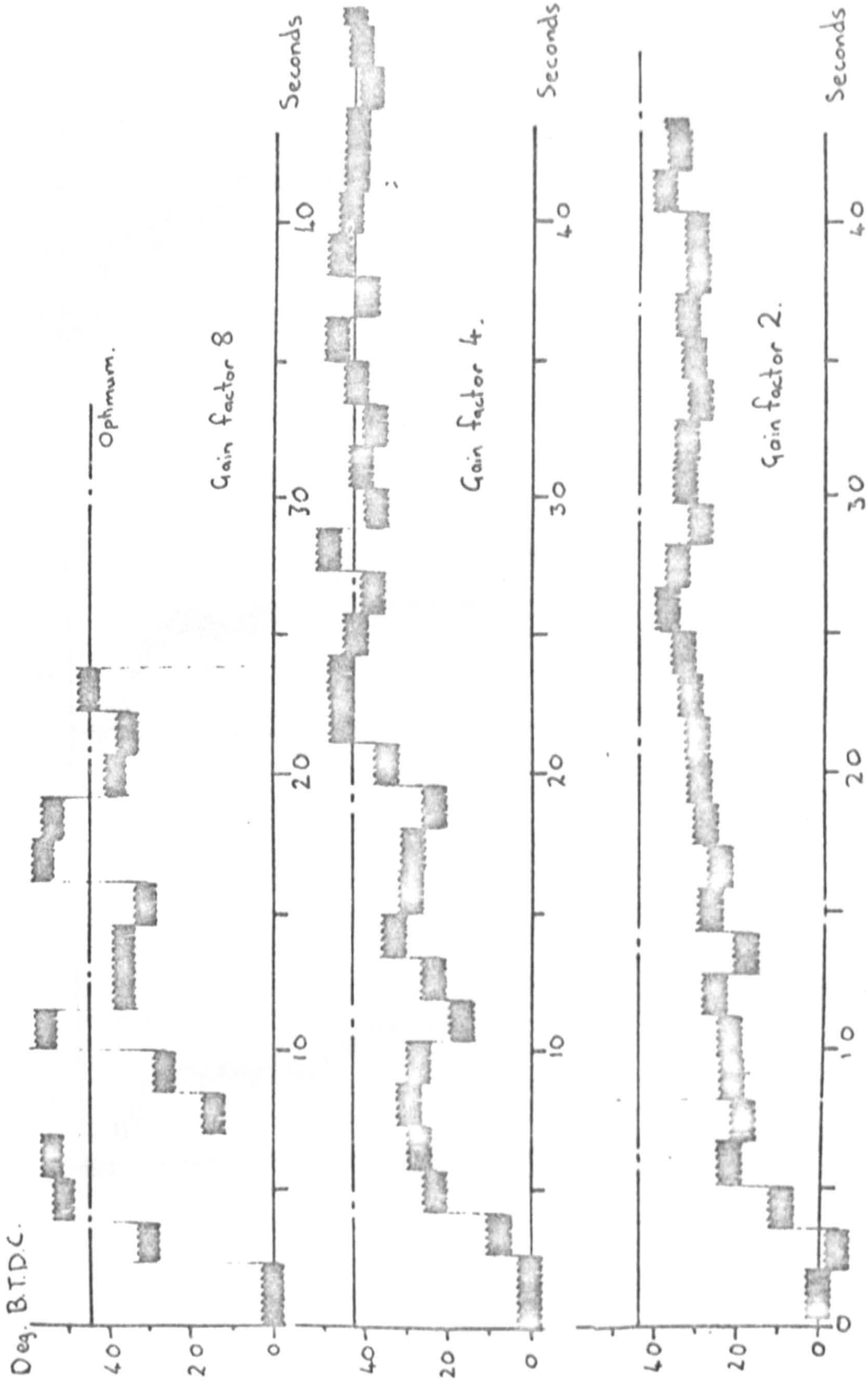
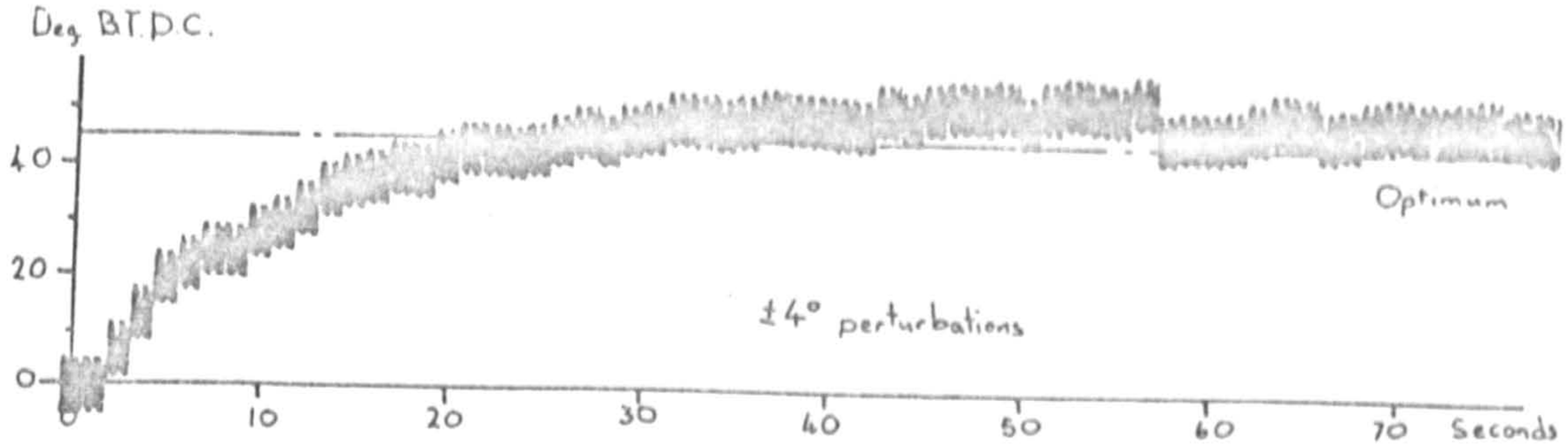
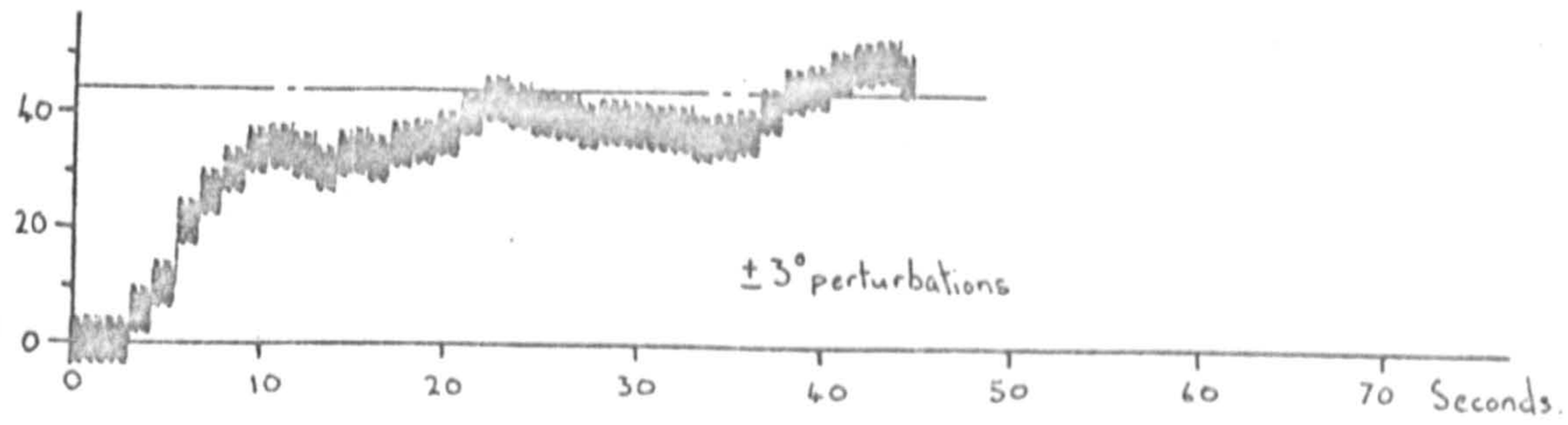


Fig. A4.1.1 Effect of optimiser gain





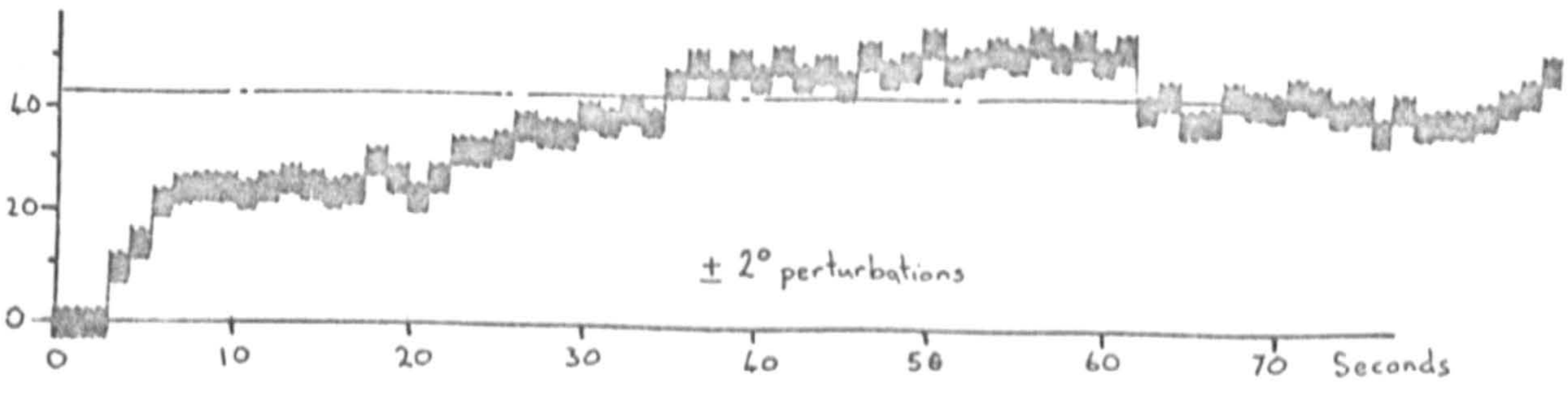


Fig. A4.1.2 Effect of perturbation amplitude

Fig. A4.1.3 Effect of allowing for longer settling times

The effects of steady state level may be removed from the impulse response estimates by assuming the system has settled within the perturbation period and the later ordinates are zero. These experiments show the effect of assuming longer settling time than necessary. The variance of the cost function slope estimates increases as fewer impulse response elements are assumed to have settled.

Perturbation clock rate ·04 sec. Perturbation amplitude ± 4^{o}

Gain factor 4

Fig. A4.1.4 Effect of averaging over several periods

These experiments show that the variance can be reduced by averaging the cost function slope estimates obtained over several periods of perturbations.

Perturbation clock rate ·04 sec. Perturbation amplitude ± 2^{o}

Gain factor 4

The impulse response was assumed to have settled in 15 clock intervals

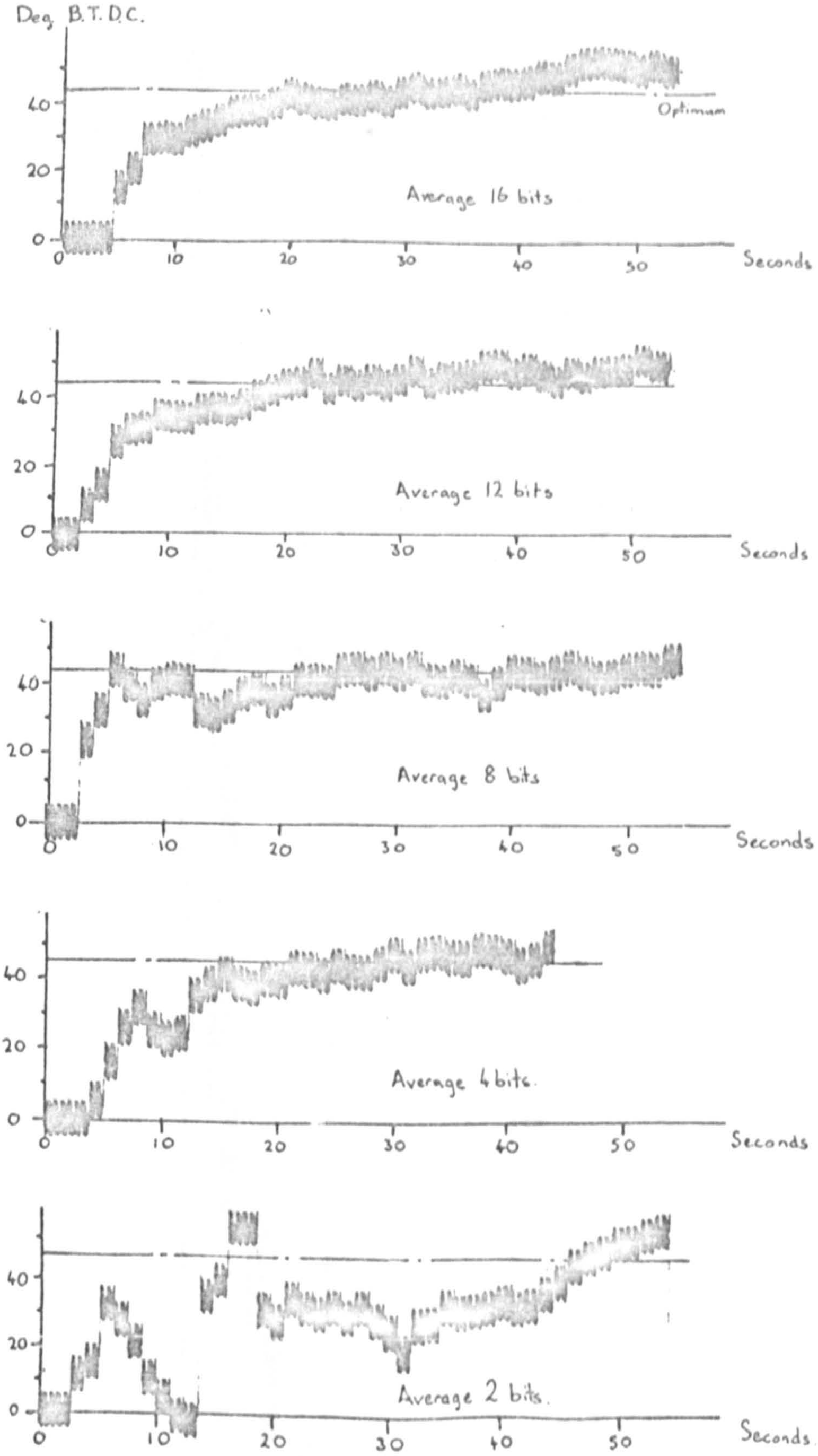


Fig. A4.1.3 Effect of allowing for longer settling times

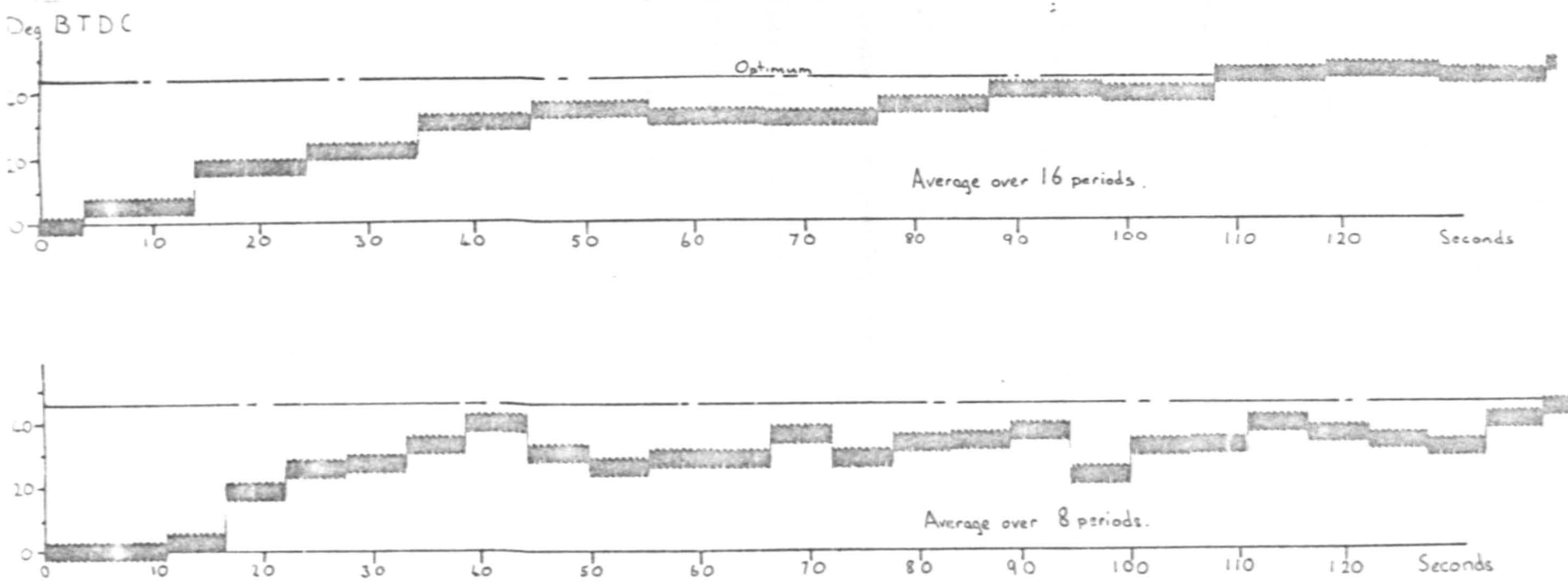


Fig. A4.1.4 Effect of averaging over several periods

Fig. A4.1.5 Effect of increasing the perturbation period

An alternative method of reducing the variance is to sample at the same rate but increment the perturbation less frequently. The increased number of samples gives a reduction in variance.

Perturbation clock rate ·16 sec. Perturbation amplitude ± 2°

System output samples ·04 sec.

The impulse response was assumed to have settled in 5 clock intervals

Fig. A4.1.6 Effect of increasing the perturbation period

These experiments are similar to the previous set but use longer perturbation sequence incrementing intervals

Perturbation clock rate ·32 sec. Perturbation amplitude ± 2°

System output sampled ·04 sec.

The impulse response was assumed to have settled in 3 clock intervals

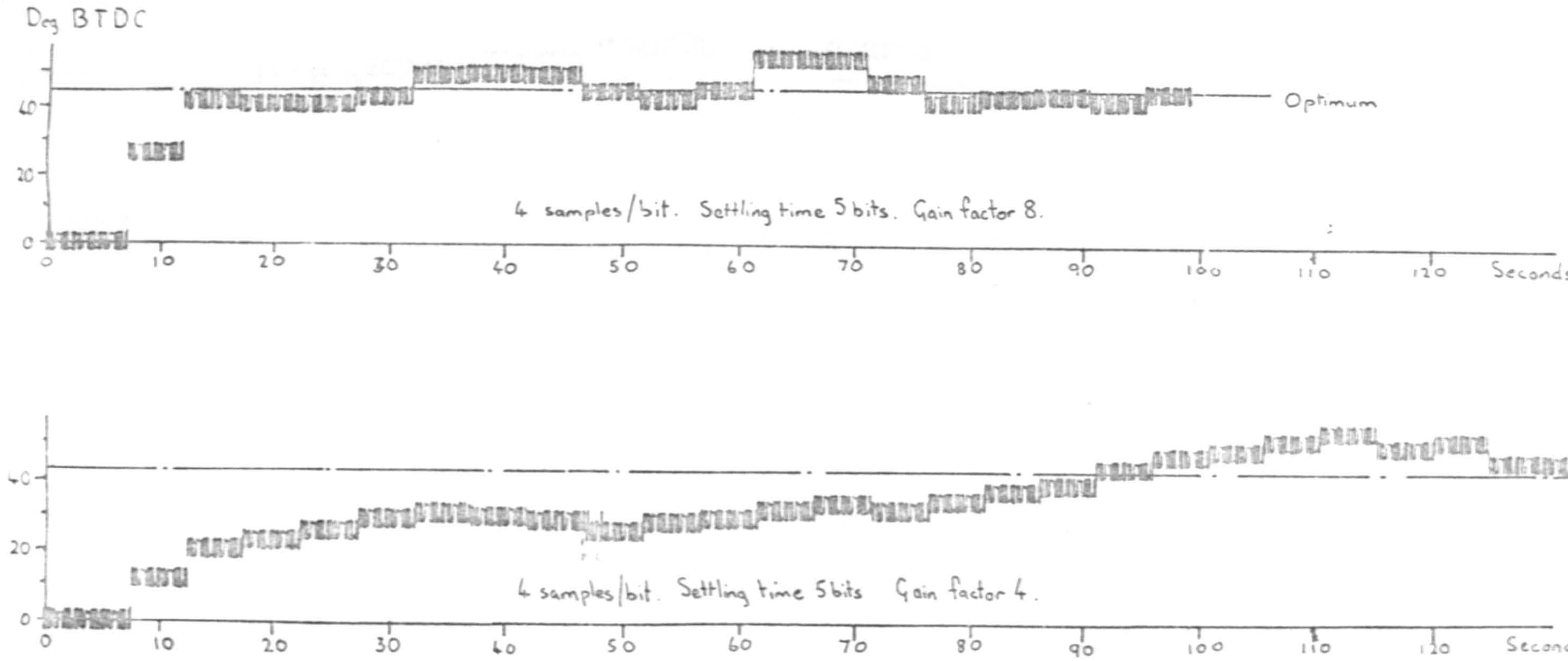


Fig. A4.1.5 Effect of increasing the perturbation period

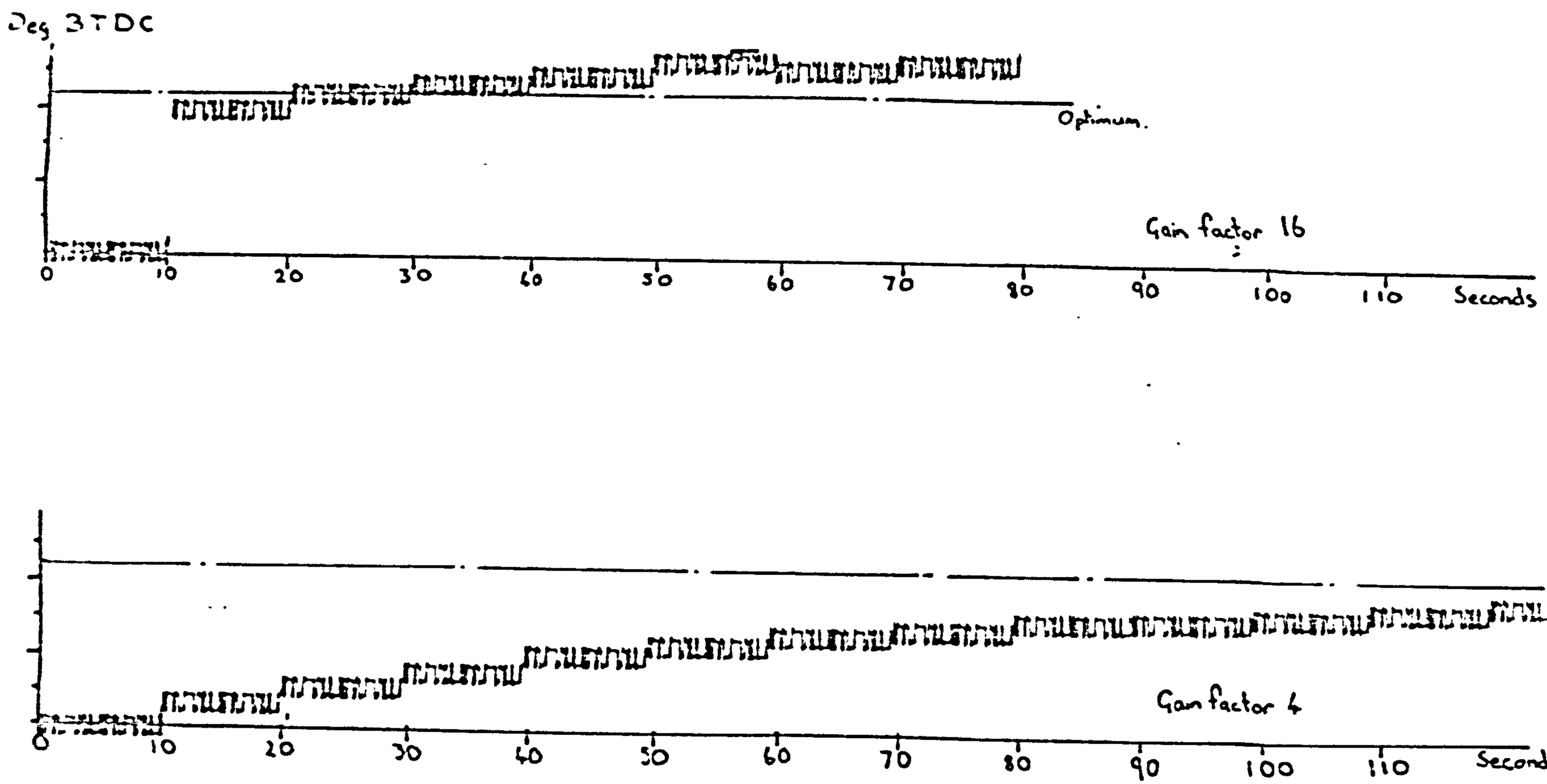


Fig. A4.1.6 Effect of increasing the perturbation period

A4.2 Pseudo-random ternary sequences

These experiments were carried out using a pseudo-random ternary sequence of length twenty-six and the engine speed controller set point was maintained at 2500 r.p.m. The system impulse response was assumed to have settled in half the pseudo-random ternary sequence perturbation period in all cases.

Fig. A4.2.1 Effect of gain factor

This shows the increase of *wander* in the optimiser as the gain is increased.

Perturbation clock rate ·04 sec. Perturbation excursion ± 3^{o}

Fig. A4.2.2 Effect of perturbation amplitude

These experiments show that the variance of the estimate of the cost function slope may be reduced by increasing the amplitude of the perturbation. Note that when the experiment used a perturbation excursion of ± 1^{o}, the noise always drove the optimiser to a limit causing the engine to stall.

The experiment was repeated with a smaller gain factor to reduce the effects of noise.

Perturbation clock rate ·04 sec. Gain factor 1.

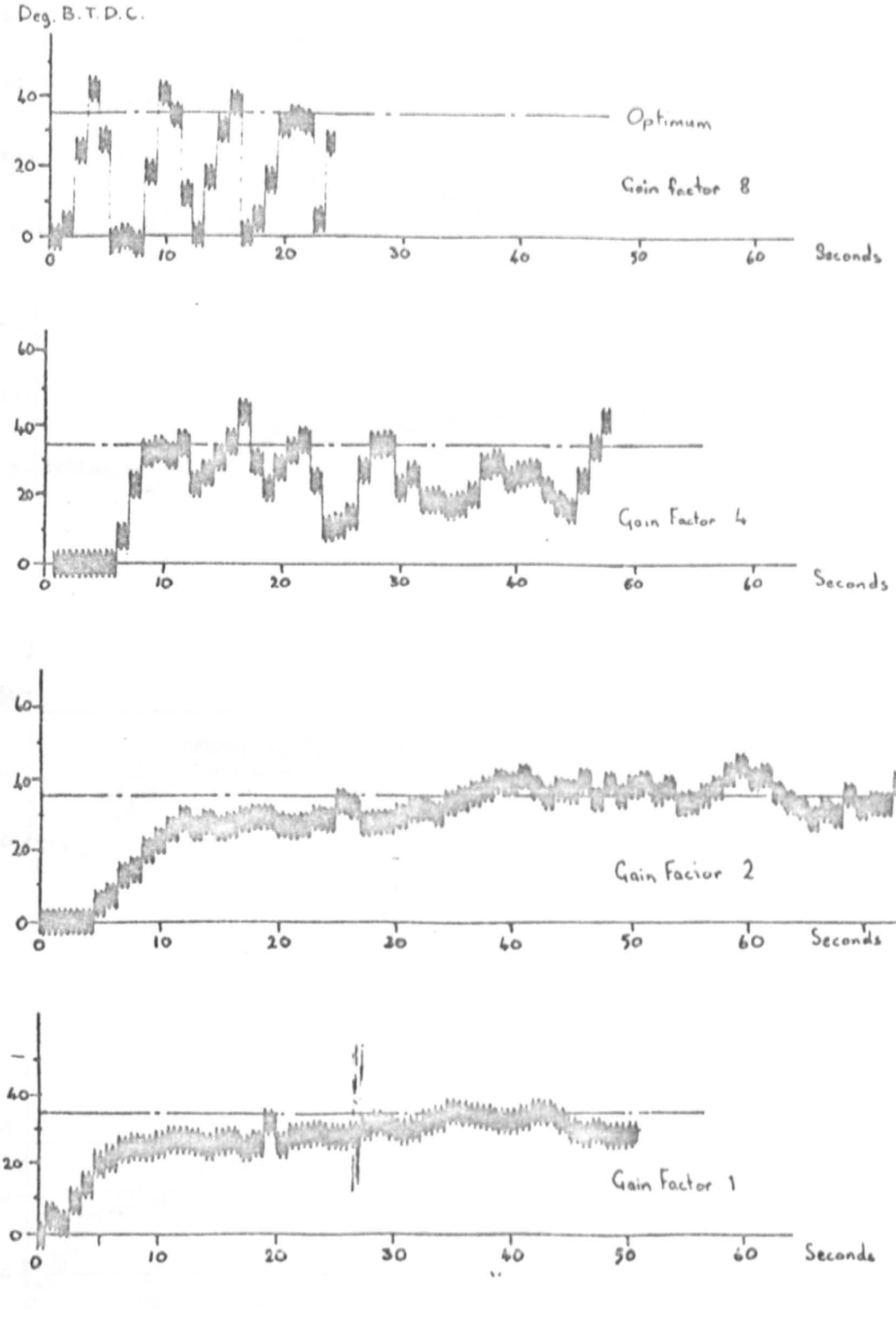


Fig. A4.2.1 Effect of gain factor

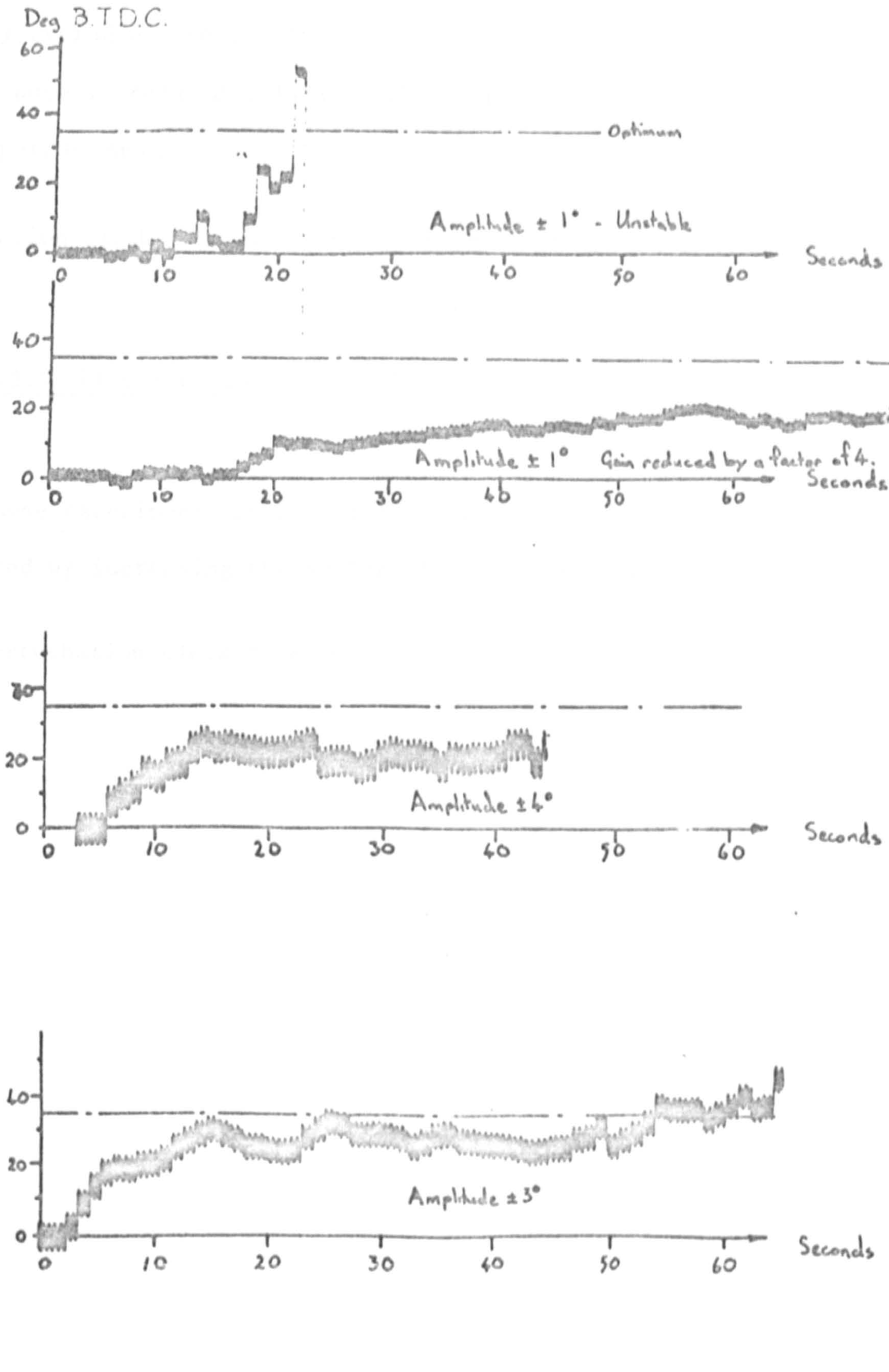


Fig. A4.2.2 Effect of perturbation amplitude

Fig. A4.2.3 Effect of averaging over several periods

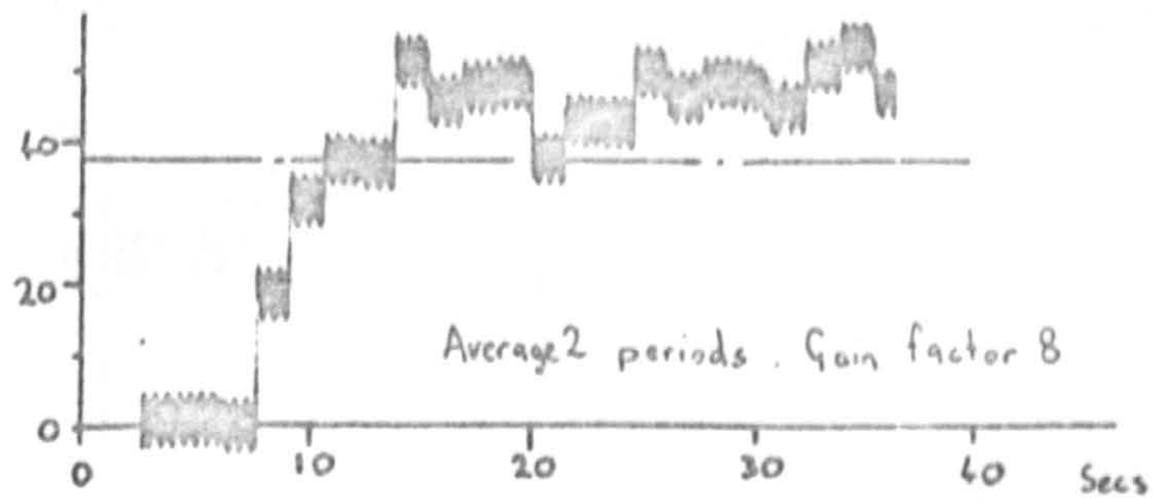
By averaging the results over several periods of perturbation, the variance is reduced and this allows higher loop gains giving a faster performance.

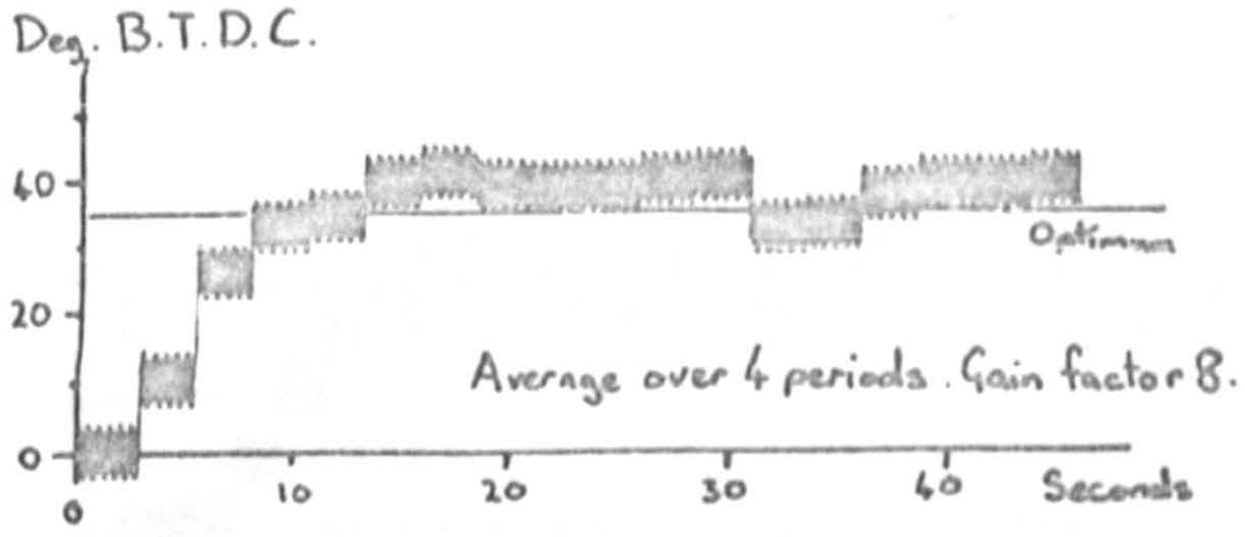
Perturbation clock rate ·04 sec. Perturbation excursion ± 3°

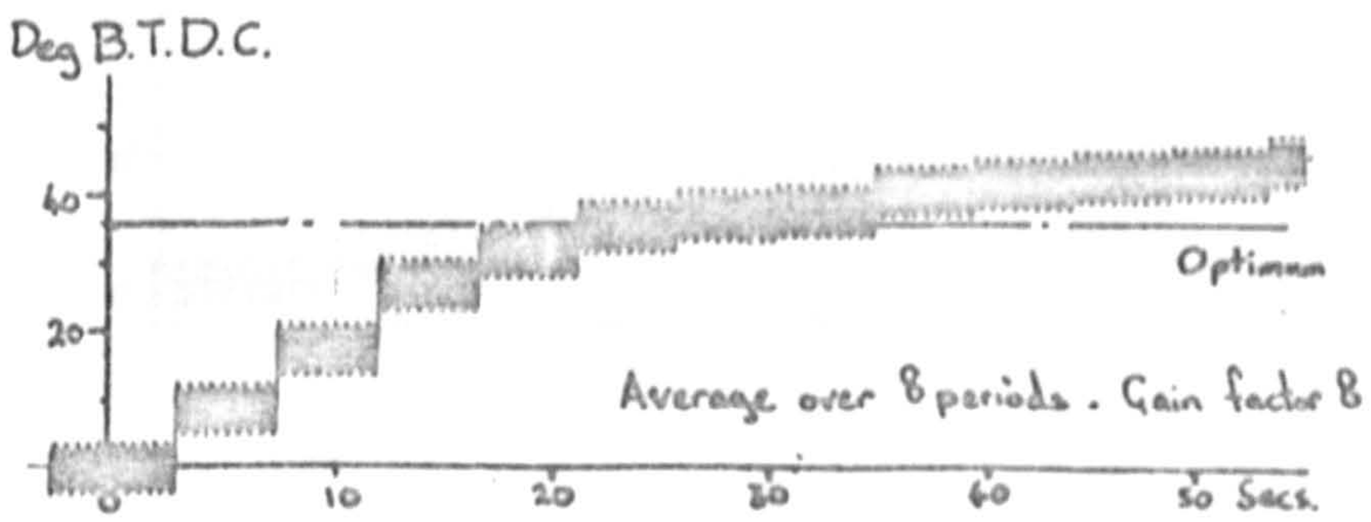
Fig. A4.2.4 Effect of increasing the period length

These experiments show that the variance of the estimate can be reduced by increasing the number of samples per period.

Basic perturbation clock rate ·04 sec. Perturbation excursion ± 3°







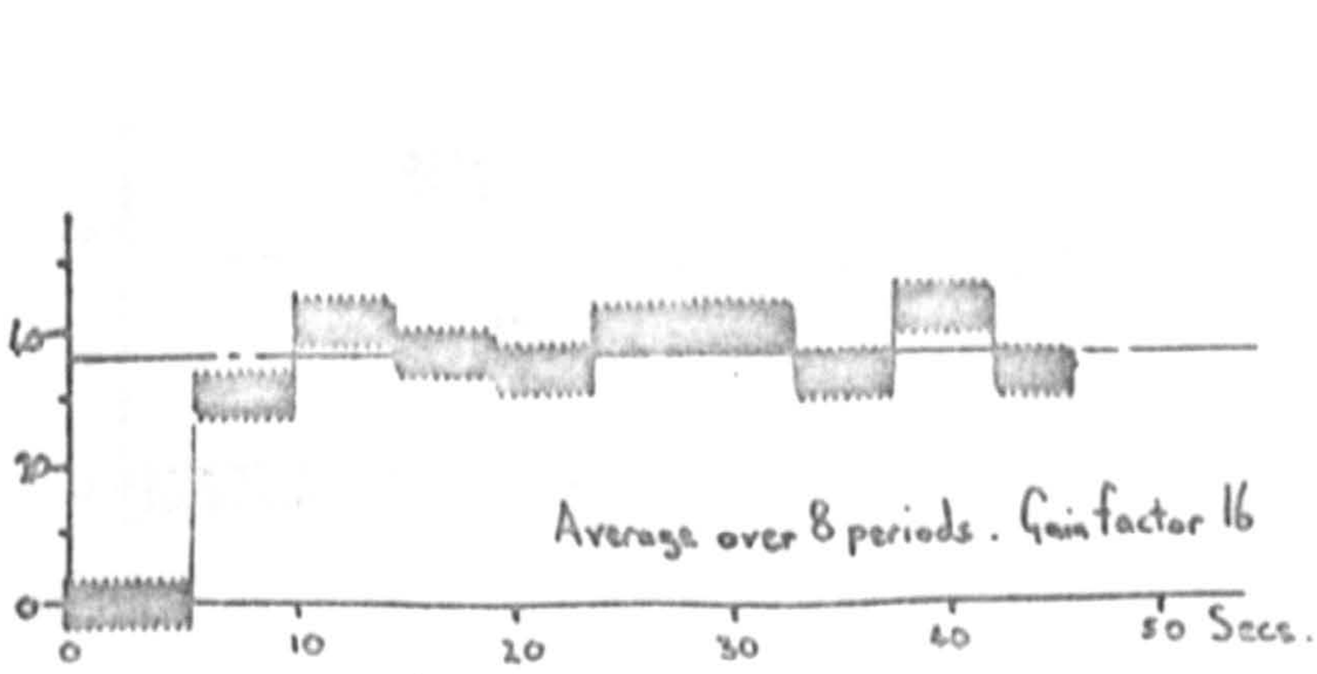


Fig. A4.2.3 Effect of averaging over several periods

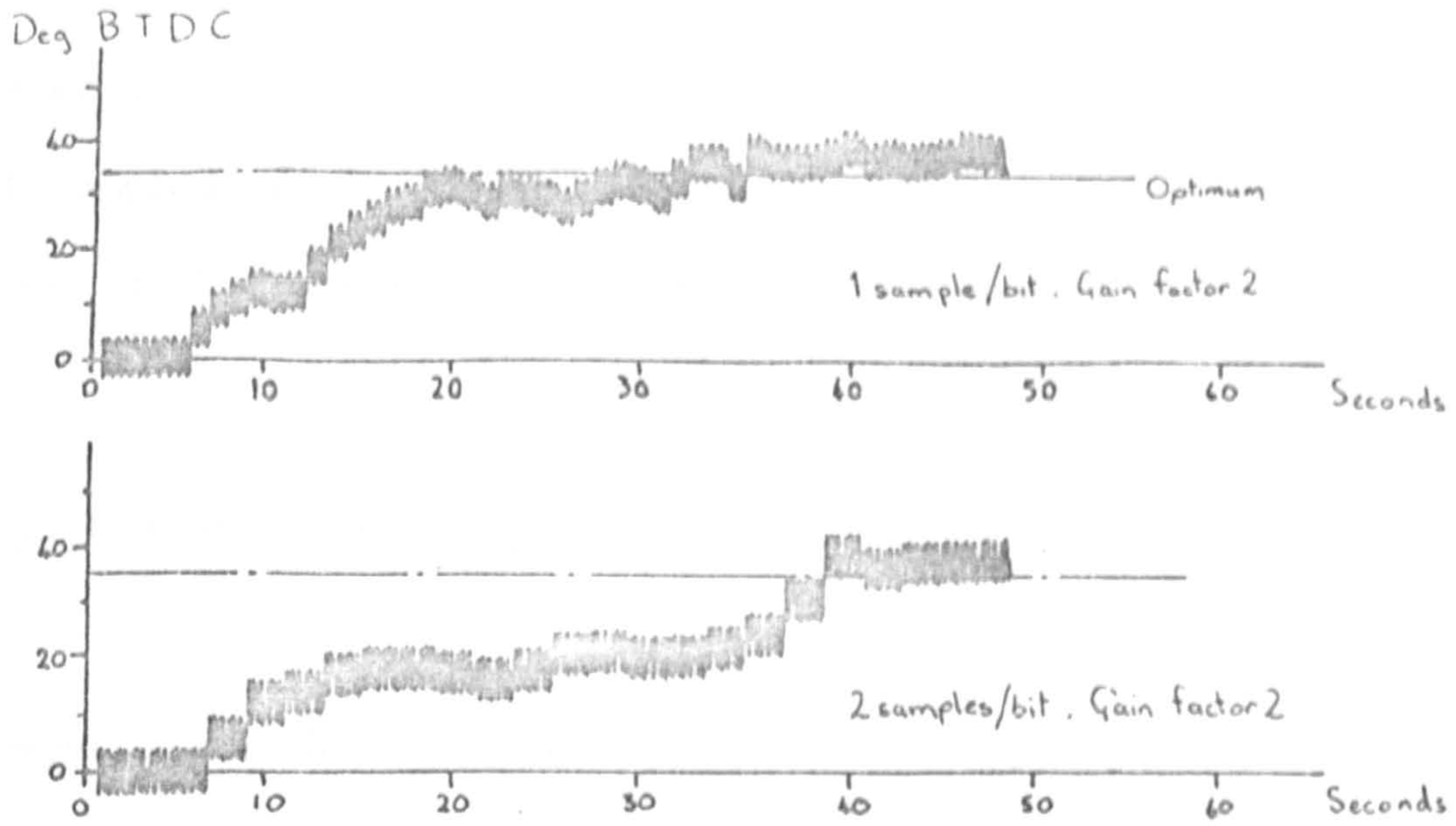


Fig. A4.2.4 Effect of increasing the period length

A4.3 Square wave perturbations

The engine speed set point was maintained at 2500 r.p.m., the system output sampled every ·04 sec. and the impulse response assumed to have settled in 15 clock intervals through these experiments.

Fig. A4.3.1 Effect of optimiser gain

These experiments show that the *wander* of the optimiser increases as the gain factor is increased.

Perturbation amplitude ± 2^{o} Total period 64 clock intervals

Fig. A4.3.2 Effect of optimiser gain with increased amplitude

These experiments are similar to the previous set except that the perturbation amplitude was increased to ± 3^{o} This reduces the variance of the estimate.

Perturbation amplitude ± 3^{o} Total period 64 clock intervals

Fig. A4.3.3 Effect of optimiser gain for longer periods of perturbation

The variance of the estimate may be reduced by assuming the same settling time but using a longer perturbation since more data is then available for the estimation of the cost function slope.

Perturbation amplitude ± 3^{o} Total period 128 clock intervals

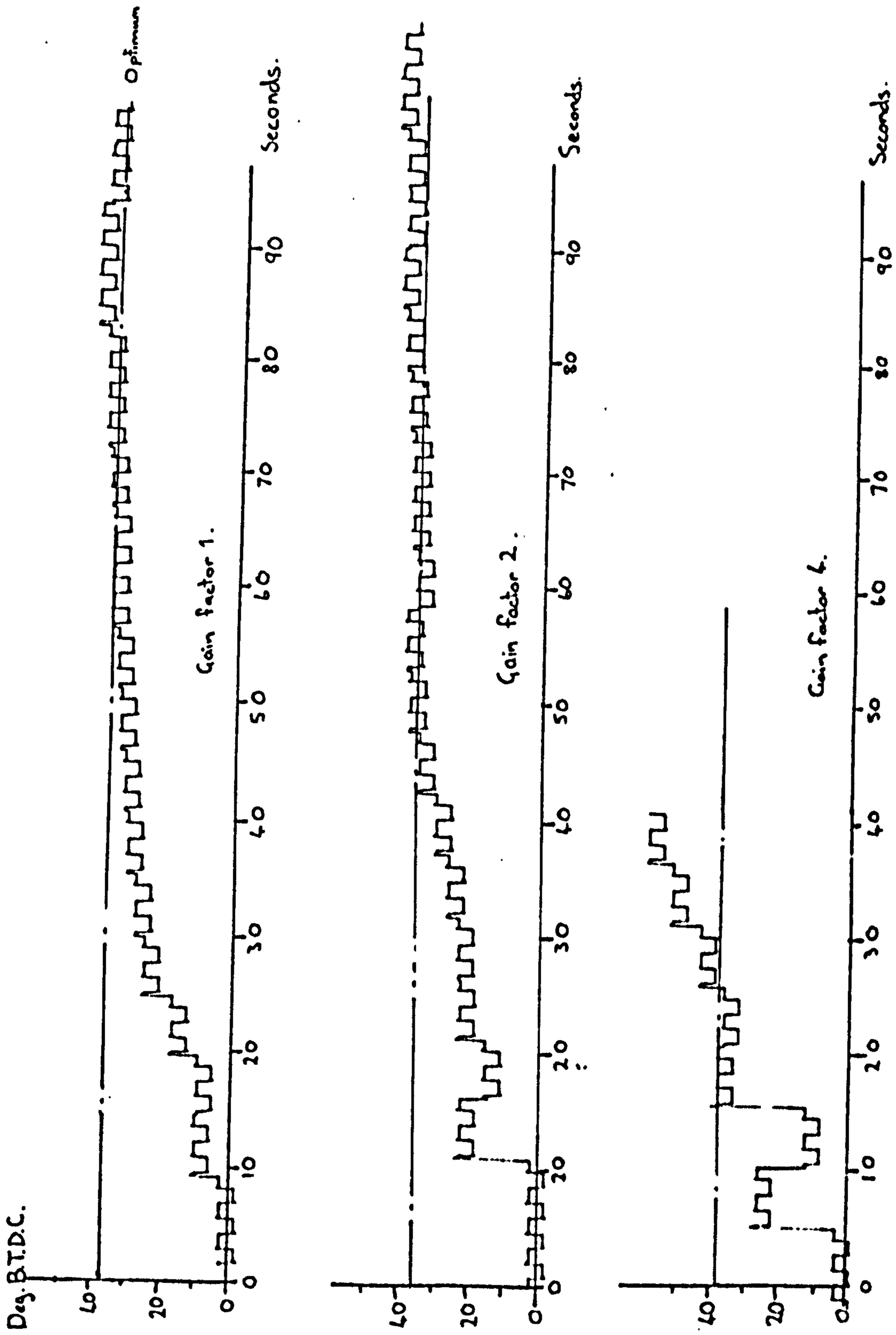
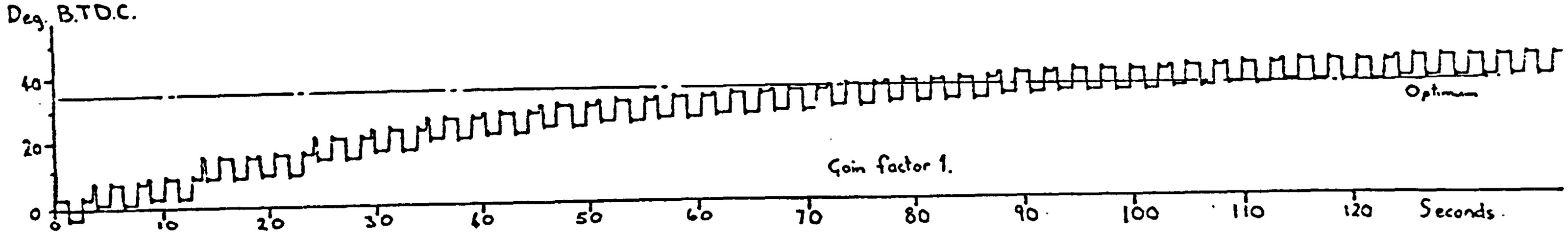
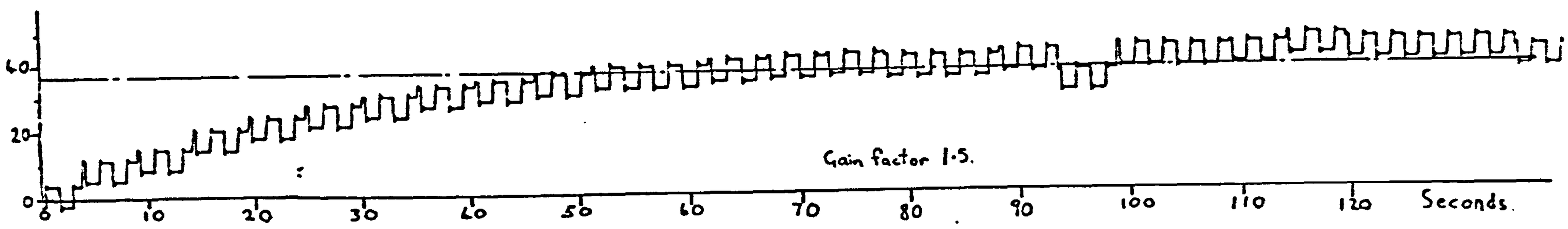


Fig. A4.3.1 Effect of optimiser gain





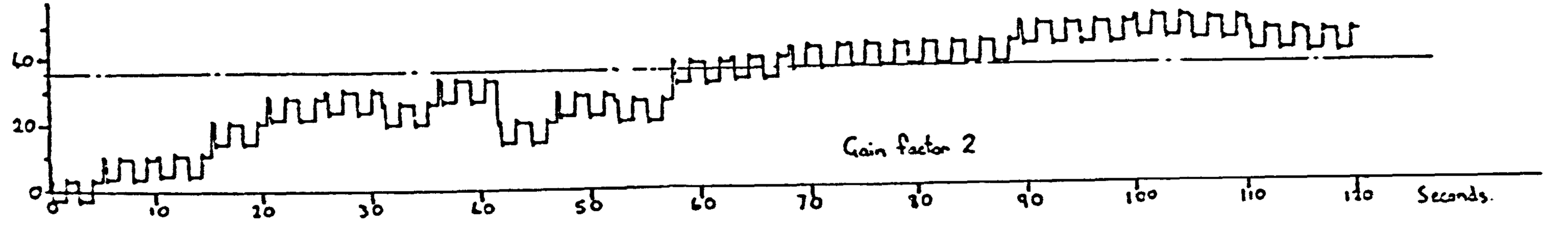


Fig. A4.3.2 Effect of optimiser gain with increased amplitude

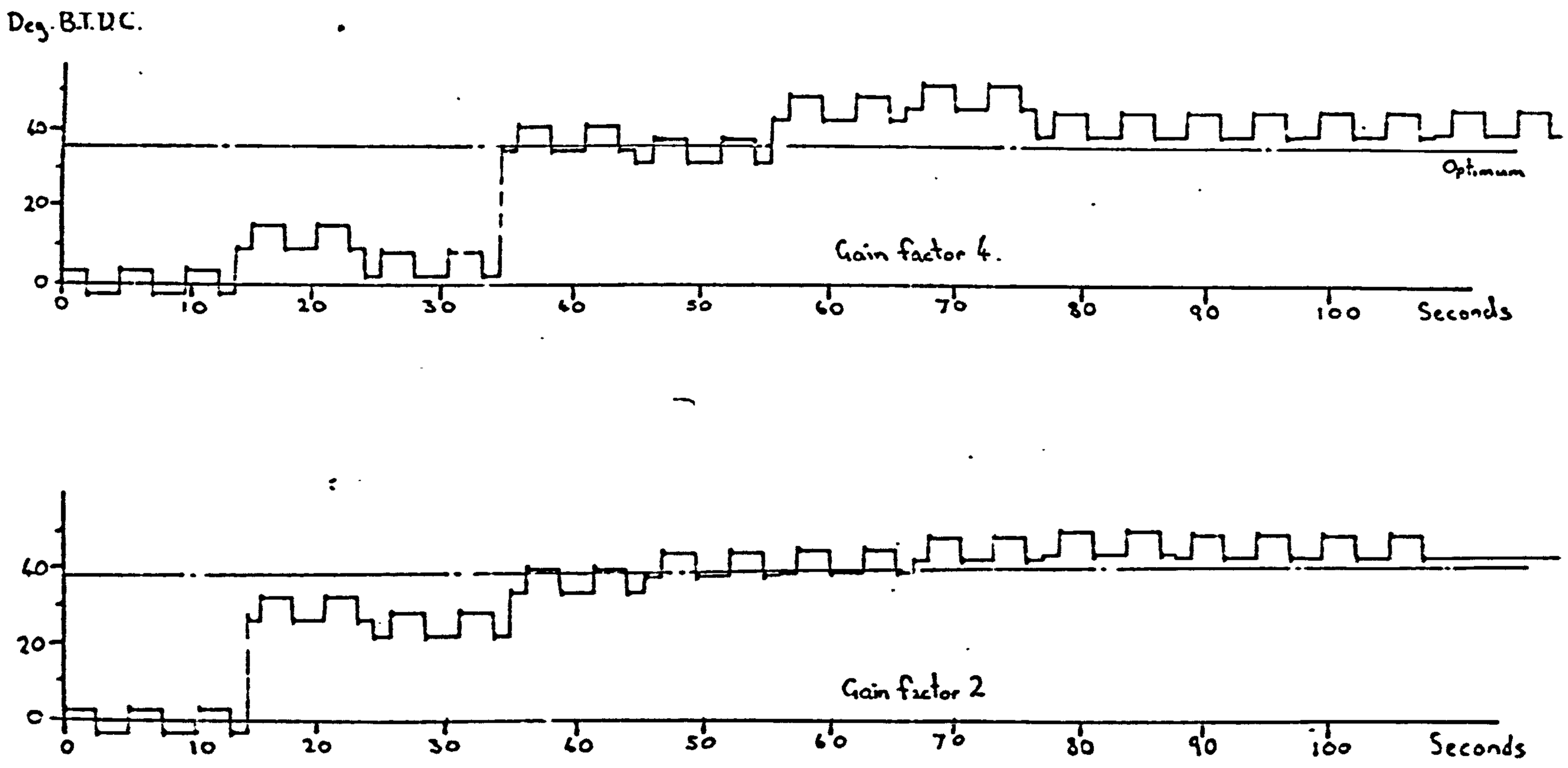


Fig. A4.3.3 Effect of optimiser gain for longer periods of perturbation